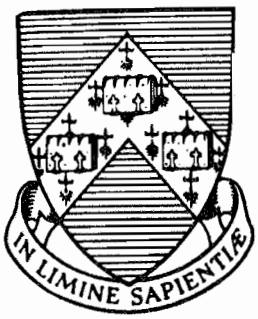


UNIVERSITY
OF YORK

CENTRE FOR HEALTH ECONOMICS

HEALTH ECONOMICS CONSORTIUM

Econometric Studies in Health Economics : A Survey of the British Literature

by

ADAM WAGSTAFF

DISCUSSION PAPER 32

ECONOMETRIC STUDIES IN HEALTH ECONOMICS:

A Survey of the British Literature

by

Adam Wagstaff

Copyright : A. Wagstaff

The Author

Adam Wagstaff was formerly Research Fellow at the Centre for Health Economics and is now Lecturer in Economics in the School of Social Sciences at the University of Sussex and a Visiting Research Fellow at the Centre for Health Economics, University of York.

Acknowledgements

An earlier version of this paper was presented to the seventh meeting of the Nordic Health Economists Study Group in Odense in September 1986. I am grateful to several participants for helpful comments, but especially to Kjeld Moller Pedersen who acted as discussant. Financial support from the Nuffield Provincial Hospitals Trust is also gratefully acknowledged.

Further Copies

Further copies of this document are available (at price £4.00 to cover costs of publication, postage and packing) from:

The Secretary
Centre for Health Economics
University of York
Heslington
York YO1 5DD

Please make cheques payable to the University of York. Details of other Discussion Papers can be obtained from the same address, or telephone York (0904) 430000, extension 5751/2.

The Centre for Health Economics is a Designated Research Centre of the Economics and Social Research Council and the Department of Health and Social Security.

ABSTRACT

This paper provides a survey of British applied econometric work in the field of health economics. The literature is divided into five main areas: the supply of health care; the demand for health care; non-medical influences on health; market equilibrium and non-price rationing; and planning, budgeting and monitoring mechanisms. In addition to surveying the literature to date, the paper also offers some suggestions for future research.

CONTENTS

1. Introduction 1

2. The supply of health care 3

2.1. Theoretical considerations on the supply side 3

2.2. Production function studies 5
2.2.1. Factor substitution and allocative efficiency 7
2.2.2. Technical efficiency 8
2.2.3. Economies of scale 9

2.3. Cost function studies 10
2.3.1. Effects of casemix on average costs 12
2.3.2. Short-run average and marginal costs 12
2.3.3. Economies of scale 13
2.3.4. Factor substitution and allocative efficiency 15
2.3.5. Economic efficiency 15

2.4. NHS factor input markets 16
2.4.1. Denmand-side issues 17
2.4.2. Supply-side issues 19

3. The demand for health care 19

3.1. Theoretical considerations on the demand side 19

3.2. Empirical studies of the demand for health care 20
3.2.1. Price elasticity of demand 20
3.2.2. Income elasticity of demand 22
3.2.3. Effect of availability 22
3.2.4. Other determinants of demand 23

4. Non-medical influences on health 23

4.1. Health production functions 23

4.2. The demand for health 24

5. Market equilibrium and non-price rationing 24

5.1. Theoretical considerations concerning market equilibrium 24

5.2. Empirical studies of market equilibrium 30

6. Planning, budgeting and monitoring mechanisms 32

6.1. Econometric models of the health care sector 32

6.2. Applications of linear programming 34

7. Whither now? 35

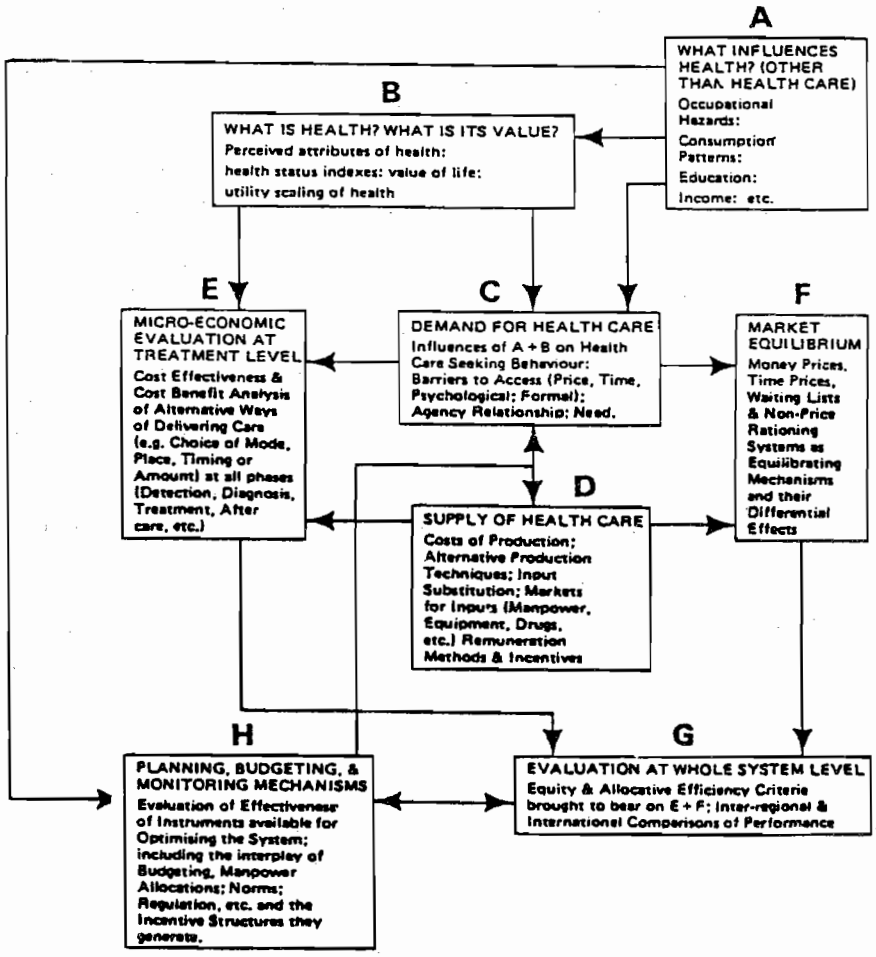
1. INTRODUCTION

Over the course of the last fifteen years or so the British literature on health economics has grown rapidly.[1] The great majority of this literature is directed at the evaluation of health care technologies: it includes a long list of applied studies in the field of economic appraisal, as well as extensive literature on the measurement and valuation of health.[2] Other areas of health economics - such as the demand for health care - have received far less attention from British researchers. It is the literature on these areas that is the subject of this survey.[3]

Much of the applied work in health economics outside the field of economic appraisal has involved the use of econometric and other quantitative methods. The earliest study of this genre was Feldstein's (1968) Economic Analysis For Health Service Efficiency, hailed by one reviewer at the time as "the best study of health services ever written by an economist" (Fuchs, 1969, p. 242). Since the publication of Feldstein's volume econometric work on the NHS has been sporadic and less extensive than might have been expected. That Feldstein's volume is still undoubtedly the 'jewel in the crown' of the British literature is a reflection of both its quality and of the paucity of high quality work that has appeared during the last twenty years. One of the objectives of this paper, therefore, is to provide an indication of the opportunities that exist for future work in the area.

The chart in Figure 1 provides a useful framework for organizing the present survey.

Figure 1: A schematic view of health economics



The British health economics literature to date has been focused firmly on boxes B and E. This survey is confined to the literature in the other boxes and begins with box D on the supply of health care.[4] Here the institutional peculiarities of the NHS mean that there is only limited scopt for useful 'importation' of studies from countries with other types of health care delivery system. It is these institutional details that determine the nature of both the questions that are likely to be of policy relevance and the problems that are likely to be encountered in exploring them. The next box covered is box C concerning the demand for health care: here institutional details play a smaller role in determining what issues

are likely to warrant research, though they are not altogether unimportant. Box A is the next box covered and includes the literature on 'health production functions' and the 'demand for health': neither, however, has been the subject of much research using British data. The next box covered is box F. Here again it is the institutional peculiarities of the NHS which provide the pointers for research activity, the almost complete absence of money prices in the NHS prompting the obvious questions: what non-price rationing devices operate? And: do they clear the 'market' for health care? Box H is the final box covered in the present survey; the literature here includes sector-wide models of the NHS and applications of linear programming to health care planning. The final section of the paper - section 7 - offers some suggestions for future research.

2. THE SUPPLY OF HEALTH CARE

The environment in which health care providers in the NHS operate differs considerably from the environment of the industrial sector or indeed that of the private health care sector. NHS providers typically do not sell their output on a pro rata basis; most of their revenue comes in the form of a grant funded mainly out of general taxation. Health care providers in the NHS also enjoy a quasi-monopoly status. Moreover, they generally do not hold property rights in any cost savings they generate. All this means that providers in the NHS tend to have more in common with the Government 'bureau' of Niskanen (1971) than with either the neoclassical firm of the economics textbook or the models of health care institutions developed for the US private health care sector (cf, Spicer, 1982).

The 'bureaucratic' nature of the NHS raises a whole range of efficiency-related issues for the funding agency: will providers be allocatively efficient? Will they be technically efficient? Will they produce the mix of outputs society values most? For the economist the NHS's 'bureaucratic' structure means that there is plenty of scope for applying the methods of econometrics to try to determine the extent of inefficiency, but also to try to provide information that may help to reduce it. However, it also means that economic concepts and econometric methods developed to analyse other sectors (notably industry) may need some modification before they can be brought to bear on the NHS.

The purpose of this section is to provide a survey of the empirical literature on the supply of health care within the NHS. Though the focus is on empirical work, it is useful to begin with a brief survey of the theoretical literature on provider behaviour that is of relevance to the NHS. This literature is not extensive and has only been partly successful in exploring the behavioural implications of the special features of the NHS environment. The paper then does on to survey the empirical literature on provider behaviour. Virtually all the studies to date have involved the use of production and cost functions; there have been no serious attempts to test behavioural models on NHS data. Later in this section the literature on NHS factor input markets is surveyed.

2.1. THEORETICAL CONSIDERATIONS ON THE SUPPLY SIDE

Only a few theoretical models of provider behaviour relevant to the NHS have been developed to data; these are listed in Table 1 and relate exclusively to the hospital sector.[5] On the issue of allocative efficiency the literature says very little: the implications of Spicer's arguments are

Author	Details of model	Predictions	Remarks
1. Feldstein (1967)	Hospital maximizes utility function (defined over cases treated, length of stay and 'quality' of care) subject to bed-occupancy constraint (cases x length of stay = 365 x beds x occupancy rate) and budget constraint (expenditure = cases x length of stay x average cost per patient day). Assumed that average cost per patient day is Cobb-Douglas function of length of stay and 'quality' of care, and that utility function is separable by logarithmic transformation.	Increase in budget (with constant bed stock) results in increase in number of cases treated and decrease in average length of stay if elasticity of utility with respect to 'quality' is decreasing function of 'quality'. Increase in stock of beds (accompanied by less-than-proportional increase in budget) results in increase in average length of stay, but a proportionally larger increase in number of cases treated, if certain "plausible" restrictions are placed on the utility function (see Feldstein, 1967, pp. 213-5).	Decision-maker not identified (the terms 'management' and 'decision-maker are used interchangeably).
2. Frost (1977)	Model of clinician behaviour; seeks to explain clinicians' assignment of patients to different treatment regimens. Clinicians form a 'prior' probability that patient X has condition Y and a 'posterior' probability following diagnostic tests. In assigning patients to treatment regimens clincians compare their expected utilities associated with alternative regimens.	The higher is a surgeon's prior probability that patient X has condition requiring surgery, the greater the likelihood of the patient receiving surgery.	Frost suggests that clinicians may alter prior probabilities following changes in available capacity. This argument - which forms the bases of the three 'hypotheses' advanced by Frost and Francis (1979) - is rather lame. Ideally model would take into account that switch-point between regimens depends not only on prior probabilities but also on hospital capacity.
3. Lindsay (1980)	Based on Lindsay's (1976) model of government enterprise. From patient's point of view hospital output comprises various 'characteristics'; some are 'visible' (eg, receipt of treatment) but others are 'invisible' (eg, provision of information, reassurance and comfort). Because invisible characteristics are impossible or costly to monitor, funding agency relies on visible characteristics when monitoring hospital's performance.	Managers of NHS hospitals will reduce cost per patient day below its 'efficient' level by cutting out expenditure on invisible characteristics. Management will encourage physicians to keep patients in hospital for longer than 'necessary' to keep cost per patient day low.	No comparative static analysis undertaken; comparative institution analysis arguably of less interest from point of view of analysing behaviour of NHS hospitals.
4. Spicer (1982)	No formal model developed. Suggestion that NHS hospitals possess many of the features of the government 'bureau' of Niskanen (1971): output not sold on pro rata basis; employees have no property rights in cost-savings; NHS hospitals enjoy quasi-monopoly status.	Incentives for NHS hospitals to produce output mix most valued by society at minimum cost very weak. Difficult for funding agency to monitor efficiency of hospitals: no prices to use as indicator of how 'outputs' are valued and no way of establishing how far costs are in excess of minimum.	

that there may be some overemployment of factors providing positive utility to decision-makers (cf, Migue and Belanger, 1974) and may even result in providers operating in the uneconomic region of their isoquent maps (cf, Gravelle and Rees, 1981, p. 166). Technical efficiency would seem to be implied by Feldstein's model, since it is an automatic corrolary of utility-maximizing behaviour (providing utility is increasing in output; cf, Stigler, 1976). Whether or not technical efficiency is implied by Spicer's arguments is not clear: lack of incentives to minimize costs may presumably result in technical inefficiency. Finally, with regard to output mix, the arguments of Lindsay and Spicer suggest that providers will definitely not produce the output mix society values most; according to Lindsay there will be a systematic bias against 'invisible' outputs, such as provision of comfort and reassurance to those undergoing treatment.

2.2. PRODUCTION FUNCTION STUDIES

Three issues in particular have dominated the British literature on production function analysis in the field of health care. The first is the issue of factor substitution. The type of questions of interest here are: to what extent are nurses substitutable for physicians in the production of health care? If they are substitutable, do considerations of allocative efficiency suggest that some substitution away from doctors towards nurses would be desirable? The second issue concerns technical efficiency. The type of questions of interest here are: is the level of technical efficiency of hospital A higher than that of hospital B? How far is the average level of technical efficiency in NHS hospitals below its feasible maximum? The third issue is that of economies of scale. Are there economies of scale in the production of hospital care? If so, up to what size of hospital?

The production function studies that have been undertaken to date on British data are listed in Table 2.6. The focus of this literature has been almost exclusively on the hospital sector; Gray's (1982) study of dental care is the only study of the primary care sector to date. The studies listed in Table 2 differ considerably in their methodology. One key respect in which they differ is in their choice of output measure. Ideally what one would like is some measure of 'value added' (in terms of health improvements) along the lines suggested by Culyer et al (1971). Empirical work is this area is, however, still in its infancy (see e.g., Williams, 1985; Gudex, 1986) and researchers have had to fall back on measures of 'throughput' as a proxy for output (e.g., cases treated, inpatient days). In instances where cases are relatively homogeneous (maternity cases, for example) this is probably satisfactory. Where, however, cases are not homogeneous (as in an acute hospital or a dental practice), the implicit assumption of assigning equal weights (or valuations) to each case becomes far less attractive.

Feldstein's (1967) solution was to divide patients into groups (or 'casemix categories') according to the department into which they were admitted and then take into account inter-hospital variations in casemix either by weighting each casetype by its expected average cost or by entering the vector of casemix proportions as a regressor in the production function. The first approach is based on the (somewhat heroic) assumption that cost per case can be viewed as a "first approximation" to marginal social valuation; the second in effect side-steps the problem of assigning weights to different casetypes. The two principal problems with this approach are that it does not address the problem of intra-category variations in case severity (not all patients admitted for surgery are

Table 2: Production function studies of the NHS

Author(s)	Sample	Output measure(s)	Input categories	Functional form(s)	Estimation method(s)	Remarks
1. Feldstein (1967)	Short-term, non-teaching general hospitals; England and Wales; 1960/61; n = 177.	No. cases, adjusted for casemix; weighted no. cases treated, where weights given by average cost of treating casetype.	Physicians; nurses; beds; drugs and dressings; catering; other expenditure.	Cobb-Douglas; Mixed Leontief Cobb-Douglas; Five-equation model of hospital production.	OLS; IV; OLS on data in first-differences; Indirect least squares.	Cobb-Douglas also estimated for sub-samples of hospitals grouped according to size; Varying-parameter version of Cobb-Douglas also estimated, where parameters depend on casemix.
2. Lavers and Whynes (1978)	Maternity hospitals; England 1971/72; n = 193.	No. cases.	Physicians; nurses; beds; drugs and dressings.	Cobb-Douglas; Translog.	OLS.	Translog estimated directly, with and without homogeneity restrictions imposed.
3. Gray (1982)	Dentists; Scotland 1979; n = 266.	Gross fees earned.	Dentists own time; surgery assistants; dental hygienists; ancilliary staff; no. chairs; dentist's age; size of practice.	Cobb-Douglas.	OLS.	Rationale for use of gross fees as output is that fee structure is designed to reflect time taken to complete course of treatment.
4. McGuire and Westoby (1983)	Non-teaching, mainly acute hospitals; Scotland, 1981/82; n = 28.	None required (see Comments).	Beds; medical and nursing staff; pharmacy and ancilliary staff; housekeeping.	Translog.	Zellner maximum likelihood.	Parameters estimated via factor share equations: output measure therefore not required (see eg, Johnston, 1984).
5. Wagstaff (1987)	Same sample as used by Lavers and Whynes (1978).	No. cases.	Physicians; nurses; beds; drugs and dressings.	Translog.	Maximum likelihood for composed error model.	Stochastic production frontier model, with technical efficiency assumed to be distributed as a half-normal error.

equally ill) and that with a small number of casemix categories one runs the risk of overlooking some of the inter-hospital variation in casemix Barlow, 1968; Fuchs, 1969; Lave and Lave , 1970; Tatchell, 1983). The approach suggested by Gray (1982) - in effect weighting cases by the time taken by the physician to complete the treatment - rests on the implicit (and, again, somewhat heroic) assumption that the time input of the physician (or dentist) is a good indicator of the social value of the treatment.

Another key difference in methodology lies in the choice of functional form. The two most popular to date have been the Cobb-Douglas and the translog (cf, e.g., Berndt and Christiansen, 1973). The flexibility of the translog is particularly attractive in the context of a multi-factor production function.[7] Estimation in the case of the translog is, however, far less straightforward than in the case of the Cobb-Douglas. The most popular approach elsewhere in economics is the so-called 'factor shares' method (see e.g., Johnston, 1984), which involves estimating the parameters of the production function from a system of factor shares equations, rather than directly from the production function itself. This approach is far less appealing in the context of the NHS, since it rests on the assumption of cost minimization. If cost minimization does not obtain, a factor's output elasticity will not be equal its share in total costs (even an average), so that the estimated method breaks down. Because one of the principal reasons for wishing to estimate production functions in this context is to detect departures from cost-minimization (i.e., the presence of allocative inefficiency), an estimation method that assumes cost minimization has little to commend it.[8]

The results of the studies can conveniently be discussed under three headings: (i) factor substitution and allocative efficiency; (ii) technical efficiency; and (iii) economies of scale.

2.2.1. Factor substitution and allocative efficiency

Only two of the studies listed in Table 2 have attempted to estimate Allen elasticities of substitution, namely those of Lavers and Whynes (1978) and McGuire and Westoby (1983). Comparison of the results is difficult, because the input categories used in the two studies differ. Lavers and Whynes employ three input categories: beds, nurses, and drugs and dressings. In the most general equation beds and drugs are estimated to be complementary, as are nurses and drugs, whilst beds and nurses are estimated to be substitutes. The estimated degree of complementarity between nurses and drugs is particularly high (partial elasticity of substitution = -4.3). McGuire and Westoby employ four input categories: (i) beds; (ii) medical and nursing staff; (iii) pharmacy and medical ancillary inputs; and (iv) housekeeping inputs. The results suggest that input categories (i) and (ii) are strongly complementary, as are categories (i) and (iv), and (ii) and (iii). Categories (ii) and (iv), by contrast, exhibit strong substitutability. Unfortunately, neither of the studies sheds any light on the issue of the degree of substitutability between physicians and nurses.

Several of the studies to date do consider this in the context of a discussion of allocative efficiency on the basis of his Cobb-Douglas estimates. Feldstein concluded that "too much is being spent on nurses, catering and other supplies and not enough on doctors, drugs and dressings" (Feldstein, 1967, pp. 100-101). He also estimated that if the hospital with the average annual budget were to reallocate optimally its budget for

medical staff, nursing staff, and drugs and dressings, its output would increase from 6,666 (weighted) cases to 10,323 cases. The ratio 6,666/10,323 = 65% therefore provides a measure of the degree of allocative efficiency of the average hospital. The results obtained by Lavers and Whynes for their Cobb-Douglas function are markedly different from those of Feldstein. These suggest that too much is spent on medical staff relative to both nursing staff and drugs and dressings, and that too much is spent on nursing staff relative to drugs and dressings. For any plausible estimate of the 'rental' cost of a bed their results also suggest that the doctor/bed ratio is too high and the nurse/bed ratio too low.

In the event, both sets of results ought probably not to be relied upon: there is extensive evidence suggesting that the Cobb-Douglas function is almost certainly too restrictive a functional form for health care institutions. Feldstein, for example, found that the estimated output elasticities for physicians and beds varied systematically with hospital size, suggesting that - in contrast to the assumption implicit in the Cobb-Douglas equation - these two output elasticities are not independent of the amounts of medical staff and beds used. Moreover, several of the output elasticities obtained by both Feldstein and Lavers and Whynes are very low, suggesting that part of the output associated with these inputs may derive from their effects on the productivity of other inputs (e.g., physician productivity in the case of nurses). The translog, unlike the Cobb-Douglas, allows both types of problem to be overcome. There is, therefore, a fairly strong prima facie case for preferring the translog estimates. Only one of the studies to date has attempted to discriminate between the two functional forms on statistical grounds: the result was a decisive rejection of the Cobb-Douglas (Wagstaff, 1987).

What, then, do the translog estimates suggest about allocative efficiency? The translog results of Lavers and Whynes, like their Cobb-Douglas results, suggest that there is probably some over-employment of physicians: indeed, the estimated marginal product of physicians is actually negative. A negative marginal product in the context of the NHS hospital is less implausible than in the context of, say, the manufacturing sector: the absence of any real incentive to minimize costs in the NHS may well result in over-employment of some factors to such an extent that their marginal products become negative. McGuire and Westoby do not report estimates of marginal products, but do report the results of a comparison of actual and optimal factor mixes. These results suggest that expenditure on beds and housekeeping are too high, whilst expenditure on medical and nursing staff combined it too low, as is expenditure on pharmacy and ancillary services combined.

2.2.2. Technical efficiency

The issue of technical efficiency has received far less attention to date than the issue of allocative efficiency. Feldstein was the first to investigate the issue and suggested using the residuals of the production functions as a measure of technical efficiency. Thus a hospital with a residual equal to zero was said to be of average technical efficiency, whilst hospitals with residuals which were greater (smaller) than zero were said to be of above-average (below-average) technical efficiency (Feldstein, 1967, pp. 110-115). The rationale behind this is that the output of a hospital with a residual equal to zero is exactly the output that would be expected of it on the basis of its estimated output elasticities. A hospital with a positive (negative) residual, by contrast, produces more (less) than it would have been expected to produce on the

basis of the estimated parameters of the production function. Feldstein then went on to propose an index of technical efficiency, defined as the ratio of actual output to the level of output predicted by its production function.

This approach has two shortcomings. First, it only enables a ranking of hospitals by technical efficiency: it provides no information on the absolute level of technical efficiency (i.e., distance from the 'frontier'). Second, it implicitly assumes all cross-sample variation in the error term is due to variation in efficiency. In reality, as Feldstein notes, the residuals are also likely to reflect random influences outside the hospitals control (viruses, for example), as well as 'statistical noise'. (Feldstein in fact went on to estimate a composed error model on panel data, with the error being assumed to be made up of three parts: a hospital-specific term (reflecting technical efficiency); a period-specific term; and a pure random term. No attempt was made, however, to estimate the size of the technical efficiency component.)

An alternative to Feldstein's approach that does not require panel data is the stochastic frontier production function model (Aigner et al., 1977; Meussan and Van den Broeck, 1977; cf, also Schmidt, 1986). In this approach the error is assumed to be composed of a symmetric term, capturing random shocks and statistical 'noise', and a one-sided term reflecting technical inefficiency. The one-sided part of the error term forces the organization to operate on or beneath (but not above) its frontier, which is itself stochastic. In a recent paper Wagstaff (1987) has employed the stochastic frontier model on the maternity hospital data used by Lavers and Whynes (op cit.). The translog frontier was estimated subject to five different sets of restrictions using the maximum likelihood method proposed by Greene (1982). In the event, the most general model (the translog function subject only to symmetry restrictions) was found to be the most consistent with the data. For this model there was no evidence of any technical efficiency in the sample.

The frontier approach has, however, a major disadvantage, namely that the results obtained regarding technical efficiency may well be sensitive to the choice of assumption about the distributions of the two components of the error term. This makes the panel data approach, suggested by Feldstein, particularly attractive, since assumptions about the distributions of technical inefficiency are no longer necessary (cf, Schmidt, op cit.). This would allow for the possibility that technical inefficiency may be non-random: indeed, if inefficiency is not non-random, it is hard to see how estimates of its extent could be of much use to a hospital's funding agency. Technical inefficiency does not, after all, presumably come 'out of the blue'.

2.2.3. Economies of scale

Evidence from production function studies of the extent of economies of scale in the NHS is mixed. The results from the Cobb-Douglas equations of Feldstein and Lavers and Whynes suggest that hospital production is subject to constant returns to scale. Again, however, in view of the restrictiveness of the Cobb-Douglas function, this result ought probably to be treated with some scepticism. Only one study to date (Wagstaff, op cit.) has attempted to descriminate between different sets of restrictions on the translog function: here it was found that even the relatively mild restrictions implied by homogeneity were rejected in favour of the more general model. Until more tests are undertaken along these

lines, the British production function literature in this area ought probably to be viewed as agnostic on the issue of economies of scale in health care.

2.3. COST FUNCTION STUDIES

Cost funcitons have been employed in the context of the NHS to investigate five main issues. First is the issue of the effects of casemix on average costs. The principal question here is: to what extent do cost variations reflect variations in casemix? The second issue concerns short-run marginal costs. The key question here is: do hospitals produce to the left of the minimum point of their short-run average cost curve so that marginal cost is less than average cost? The third and fourth issues are issues that have already been discussed in the previous section and concern economies of scale and factor substitution. The final issue concerns the measurement of economic efficiency (i.e., technical and allocative efficiency). The type of questions here are: is the level of economic efficiency of hospital A higher than that of hospital B? How far is the average level of economic efficiency in NHS hospitals below its maximum possible? What proportion of economic inefficiency is allocative inefficiency and what proportion is technical inefficiency?

The principal cost function studies of the NHS are listed in Table 3.[7] As in the case of the production function literature, there is considerable variation in methodology. A key difference again is in the choice of output measure. The two earliest studies (Feldstein, 1967; Hurst, 1977) both employed the second of Feldstein's approaches to measuring hospital output, namely treating the vectors of casemix proportions as a regressor. Several of the more recent studies (e.g., Culyer et al., 1978; McGuire and Williams, 1986) have employed the information theory approach proposed by Evans and Walker (1972): this provides a measure of hospital output based on the degree of case complexity. In some instances the problem of casemix adjustment has not arisen, due either to the use of a sample of single specialty hospitals or to the estimation of the translog function by the shares method.

There is even greater variation in the choice of model specification in the cost function literature; this is true not only across studies but also within studies. The majority of studies have focused on the average cost function, though some authors have opted for the total cost function. In some instances authors have estimated both functions, though the two are seldom derived from one another and are in fact often inconsistent with one another (see e.g., Feldstein, 1967; McGuire and Williams, 1986).

The inconsistencies between total and average cost functions within studies reflects the lack of consensus in health economics about the appropriate specification of a cost function for health care institutions. The observation by Evans (1970) that the absence of incentives to minimize costs means cost functions need to be interpreted as 'behavioural' rather than technological functions has tended to result in an 'anything goes' attitude to model specification. Thus, for example, some studies include the stock of beds, caseflow, bed occupancy rate and length of stay all as independent variables (probably not a very wise move, since all are interrelated via the bed occupancy constraint); others have included only subsets of these variables, but different subsets are discovered in different studies. The justification for including beds in the cost function is that it is a proxy for capacity; theoretical justifications for including other variables from the bed occupancy constraint are few and

Table 3: Cost function Studies of the NHS

Author(s)	Sample(s)	Output measure	Model(s)	Remarks
1. Feldstein (1967)	Short-term, non-teaching general hospitals; England and Wales, 1960/61; n = 177.	No. cases, adjusted for casemix via vector of casemix proportions.	Variety of specifications estimated; final average cost equation expresses average cost per case as a function of casemix, beds, beds squared, caseflow (no. cases + beds) and caseflow squared.	Total cost equation also estimated: see section 2.3.2.
2. Hurst (1977)	Non-teaching acute hospitals; non-teaching single specialty hospitals; England and Wales, 1969/70; n = 30 (acute), n = 22 (single specialty).	As Feldstein (1967) for acute hospital sample; no. cases for single specialty sample.	Average cost per case expressed as a function of casemix and length of stay.	
3. Culyer et al. (1978)	Acute teaching and non-teaching hospitals with 100 beds or more; England and Wales, 1969/70; n = 268.	No. cases; casemix adjustment via Evans-Walker information theory-based index of case complexity.	Average cost per case expressed as funtion of casemix, beds, beds squared, occupancy rate, length of stay, caseflow, London and teaching dummies.	Occupancy rate, length of stay, beds and caseflow all included, despite fact that all are interrelated via the bed-occupancy constraint (ie., no. cases x length of stay = 365 x beds x occupancy rate).
4. Culyer and Drummond (1978)	Type 2 teaching hospitals; England 1969/70; n = 38.	As Culyer et al. (1978)	As Culyer et al. (1978) except information theory-based index of specialization also added.	
5. Steele and Gray (1982)	Maternity hospitals; Scotland, 1976/77; n = 28. Sample split into (i) specialist maternity hospitals and (ii) GP-run hospital units.	No. cases (ie, deliveries), adjusted for 'difficulty' of case; no. inpatient days.	3 models: (i) total cost expressed as cubic function of no. cases; (ii) cost per case expressed as function of no. beds, occupancy rate, length of stay, nurse-inpatient ratio; (iii) cost per inpatient day modelled as in (ii).	
6. McGuire and Williams (1986)	Acute hospitals with 50 beds or more; Scotland, 1980/81; n = 49.	No. cases with casemix adjustment via Evans-Walker information theory-based index of case complexity; no. inpatient days.	Total and average cost functions estimated; independent variables include information theory-based measures of case complexity and specialization, no. beds, occupancy rate and vector of age proportions.	Specifications of total and average cost functions inconsistent with one another; models suffer from definitional singularity, since all age-category variables were included.
7. Gray et al. (1986)	Time-series, aggregate-level data on Scottish acute hospitals, 1951-81.	None required (see comments).	Translog cost function with costs as a function of output and factor prices.	Parameters estimated via factor shares equations: output measure therefore not required (see eg, Johnston, 1984).

far between. One possible justification, however, is based oin rate-volume theory: Mann and Lett (1968) have suggested that a hospital's average costs may depend not only on its volume of output, but also on the rate at which its output is produced. If the stock of beds is viewed as a proxy for the anticipated volume or output and caseflow as a proxy for the rate at which output is produced, the simultaneous inclusion of beds and caseflow may be justified. There may, of course, be other justifications: if so, they really ought to be made explicit.

The results of the studies to date are discussed under five headings: (i) effects of casemix on average costs; (ii) short-run average and marginal costs; (iii) economies of scale; (iv) factor substitution; and (v) economic efficiency.

2.3.1 Effects of casemix on average costs

Several of the studies listed in Table 3 shed light on the issue of how far inter-hospital variations in average costs can be attributed to casemix variations. Feldstein (1965, 1967) found that 27.5% of the sample variation in cost per case could be 'explained' by variations in the casemix vector, but that only 2.1% of variation in cost per patient day could be attributed to casemix variations. This led him to conclude that, contrary to what was often assumed at the time, variations in costs per case, and - to a lesser extent - costs per patient day, do reflect variations in casemix. Indeed, it seems probable that the importance of casemix variations was underestimated by Feldstein. Evans (1971), for example, using a factor analysis-based output measure found that casemix variations explain a much higher percentage of variations in costs per case in his Canadian data. In the few British studies to date using the information theory approach no estimates of the relative 'importance' of casemix variations, in terms of either 'addition to R-squared' or 'beta coefficients' (cf, eg., Pindyck and Rubinfeld, 1981, pp. 90-91).

One attraction of Feldstein's casemix adjustment method is that it provides a means of estimating average costs per case in each casemix category. Feldstein reported estimates of casetype-specific costs for 28 casemix categories based on a regression of cost per case on a constant term and a (1x27) vector of casemix proportions. Not surprisingly, perhaps, in view of the dimensions of the casemix vector, several of the estimated average costs were implausible; indeed in some instances the figures were negative. More reasonable results were obtained by Hurst (1977) using a smaller number of casemix categories: these results were later used by Culyer and Maynard (1981) in their estimates of the hospital costs associated with treating patients with duodenal ulcers by surgery.

2.3.2 Short-run average and marginal costs

One issue that has obvious policy implications in this area is where health care institutions operate on their short-run average cost curves. Do hospitals, for example, operate to the left of the point of minimum short run average cost, so that their marginal cost is lower than average cost? If they do, use of average cost data in economic appraisals will result in some unwarranted bias against hospital-based treatments. Marginal costs can be estimated either on a cost per case basis or on a cost per patient day basis. In each case a distinction can be drawn between the marginal cost associated with a fixed stock of beds and the marginal cost associated with a fixed stock of occupied beds: in the first

case both length of stay and the bed occupancy rate are allowed to change, whilst in the second case only the length of stay is permitted to change.

Feldstein estimated the various types of marginal cost from total cost equations: total cost was expressed as a function of casemix, the number of cases treated and the stock of beds (in the case of the first type of marginal cost when only the stock of beds is fixed) and occupied beds (in the case of the second type of marginal cost when the number of occupied beds is fixed). Linear versions of these equations give estimates of marginal cost on a per case basis that are equivalent to 21% of the average cost per case (in the case when only the stock of beds is fixed) and 12% of the average cost per case (when the occupancy rate is also fixed). Marginal costs on a per patient day basis can also be inferred from the parameter estimates of these equations: these are equivalent to 54% of the average cost figure (in the case where both length of stay and the occupancy rate can change) and 74% of the average cost figure (in the case where both length of stay can change). All these results are found to be robust in the face of changes in both the functional form (quadratic and cubic specifications were tried) and the estimation method (instrumental variables was used to overcome the endogeneity of the number of cases treated).

Hurst (1977) also estimated marginal costs for NHS hospitals, but did so using an average cost function rather than a total cost function. Cost per case was specified as a linear function of length of stay and a vector of casemix proportions. On the basis of this equation Hurst estimates the cost of an additional patient day (allowing only length of stay to change) at the equivalent of 51% of the average cost per patient day in acute hospitals at the time (1969/70); this is substantially lower than Feldstein's estimate (74%).

2.3.3 Economies of scale

Following Feldstein (1967) the standard approach to investigating the extent of economies of scale in hospital care has been to estimate equations relating cost per case to casemix, the stock of beds and (in some instances) caseflow. The stock of beds is seen as a measure of capacity, and the partial relationship between cost per case and the stock of beds is viewed as evidence about the extent of economies of scale.

Feldstein began with an equation which excluded caseflow, but included the square of the stock of beds, as well as the stock of beds itself. In this equation neither coefficient on the two bed variables was statistically significant; Feldstein concluded that there were no significant economies or diseconomies of scale for the hospitals in his sample. He then went on to re-estimate the equation with caseflow and its square included, finding that the coefficients on the two bed variables were now significant with the signs suggesting a U-shaped relationship between average cost and hospital size. Feldstein interpreted this as evidence of unrealized economies of scale: the apparent non-existence of economies of scale was, he argued, due to a pure scale effect (tending to reduce cost per case) being offset by a caseflow effect (tending to increase cost per case, since larger hospitals have lower caseflows). In the absence of the caseflow effect, cost per case would reach a minimum at a capacity level of 903 beds.

Broadly similar results were obtained by Culyer et. al. (1978), who regressed cost per case on the stock of beds, its square, caseflow, casemix

and various other variables: the coefficients on beds and the square of beds were significant and indicated a U-shaped relationship reaching a minimum at 430 beds. Steele and Gray (1982) obtained slightly different results. They estimated a total cost equation in which total costs were related to cases treated (used as a measure of capacity) via a cubic function. In the event the coefficients on the two terms raised to powers were insignificant and the other coefficients indicated a slowly declining average cost curve: the results therefore literally imply ever-increasing economies of scale that are realized rather than potential.

Because of the apparent importance of the caseflow effect, it is of some interest to know the source of caseflow variations. With a fixed stock of beds caseflow can be reduced either by reducing the bed occupancy rate or by increasing average length of stay. By re-estimating his first equation again, initially with the bed occupancy rate and its square included, and then with length of stay and its square included, Feldstein found that it is primarily longer lengths of stay that are responsible for the lower caseflows of larger hospitals. He suggests that this is a reflection of a lower level of managerial efficiency in larger hospitals and suggests measures aimed at increasing the intensity of capacity utilization in larger hospitals.

Feldstein's work on economies of scale has been challenged by various authors. Some have accepted his approach, but have questioned his assertion that lower caseflows necessarily reflect managerial inefficiency. A number of writers (see e.g. Barlow, 1968; Fuchs, 1969; Lave and Lave, 1970) have argued that larger hospitals have longer lengths of stay because they treat the more severe cases within each of the casemix categories. To the extent that variations in intra-category case severity are captured by the caseflow variable, the economies of scale suggested by the specification with caseflow included may well be actual, rather than merely potential.

Other authors have questioned Feldstein's approach to the issue of economies of scale, suggesting that it is inconsistent with the theory behind the long-run cost function. The estimated relationship between cost per case and the stock of beds can only properly be viewed as a long-run average cost curve if it is the envelope of all the short-run average cost curves. The cross-section sample used by Feldstein gives a 'snapshot' of observations on average cost, the stock of beds and output for a large number of hospitals of different sizes. The sample observations on average cost and the stock of beds are then assumed to represent points on a long-run average cost curve. The observed combination for a given hospital is therefore assumed to represent a point of tangency between the short-run average cost curve it happens to be on at the time (given its bed stock) and the sample's long-run average cost curve. This will only be only the case, however, if each hospital sets its output in the short-run at the level for which its current bed stock is equal to the long-run cost-minimizing bed stock; i.e., at the point of tangency between its short-run and long-run average cost curves. As Davis (1968) has noted, there is no compelling <u>a priori</u> reason for supposing that hospitals do follow such a policy of output determination. If hospitals do not choose their output levels in this way, some other assumption needs to be made about output determination in order to determine how the estimated curve relates to the true long-run cost curve. One possibility that has been suggested is that hospitals select their output in the short-run so as to produce at the lowest point on their short-run average cost curve (Mann and Yett, 1968). If this is true, the estimated relationship between average cost and beds will represent a curve connecting these minimum points; this will, as Mann

and Yett note, clearly be steeper than the true long-run curve, thus exaggerating the extent of economies of scale.

2.3.4 Factor substitution and allocative efficiency

It is well known that, under conditions of cost-minimization, the parameters of a firm's production technology can be estimated from its cost function. Under these conditions estimation of the cost function provides an alternative way of investigating the issue of factor substitution. Because all hospitals in the NHS face the same (nominal) factor prices, estimation of the cost function only provides information on factor substitution if estimated on time-series or panel data.

The only study in the British literature to date that adopts this approach is that of Gray et. al. (1986), who estimate a translog cost function on aggregate time-series data using the 'factor shares' estimation method. Their results suggest a fairly high degree of substitutionality between capital and labour, and between professional/technical staff and domestic/ancillary staff, but a fairly high degree of complementarity between medical staff and professional/technical staff. No estimates of the extent and direction of allocative inefficiency are reported. The estimated (partial) elasticities of substitution reported by Gray et al. are subject to the same caveats as those reported by McGuire and Westoby (1983), namely that they depend crucially on the validity of the assumption of cost minimization. When the cost function is estimated directly, one can argue, as Evans (1971) has, that any absence of incentives for cost-minimization in the hospital sector merely means that the estimated cost function has to be interpreted as a 'behavioural' cost function; any estimated elasticities of substitution obtained from such an equation will, as Culyer et al. (1982) note, reflect "not only technological substitution possibilities, but also those that happen to suit - for whatever reason - those with decision-making authority" (Culyer et al., op cit., p. 132). One cannot, however, resort to such an argument in the context of the 'factor shares' estimation method, since the parameter estimates are derived under the assumption that cost-minimization does indeed obtain.

2.3.5 Economic efficiency

Cost functions may also be useful in investigating the 'economic' (i.e., allocative and technical) efficiency of health care institutions. This can be pursued either in a relative sense (Is hospital A more efficient than hospital B? Has hospital A become more efficient over the last year?), or in an absolute sense (How far are the average hospital's costs in excess of their feasible minimum?). It may also be of some interest to try to break down economic inefficiency into its two components: what proportion of observed inefficiency is due to allocative inefficiency and what proportion due to technical inefficiency?

Surprisingly perhaps, in view of its obvious policy relevance, these issues have not received much attention in the literature. Feldstein (1967) proposed an index of hospital 'costliness' based on the residuals of his basic cost functions (i.e., those including only casemix proportions as independent variables). Costliness is defined as the ratio of actual cost per case for the hospital in question to the cost per case that would be expected if its cost per case in each of the casemix categories were the same as the national average. Since the coefficients on the casemix

proportions can be interpreted as estimates of casetype-specific costs, the costliness index can be defined as the ratio of actual to predicted cost per case. The index is in effect, therefore, based on the residuals: when actual and predicted cost coincide, the residual is zero and the index takes a value of one; when actual costs exceed (fall short of) predicted costs, the residual is positive (negative) and the index takes a value that is greater (smaller) than one. Feldstein computes the index for each hospital in his sample and finds that it correlates only mildly with the index of relative cost per case (the ratio of actual cost per case to the national average) and concludes that crude costs per case are not a good measure of hospital cost performance.

The 'costliness' index can be broken down into two parts in order to show the relative importance of technical and allocative inefficiency (see Feldstein, op cit., pp. 31-33): 'costliness', C^*, can be shown to be equal to $[P^*I^*]^{-1}$, where P^* is the index of technical efficiency or 'productivity' of section 2.2 and I^* is an index of allocative efficiency or the 'appropriateness of input proportions'. The reciprocal of each component of the costliness index can be interpreted in terms of excess costs: P^{*-1}, for example, can be interpreted as the proportional increase in costs over the average attributable to lower technical efficiency. Feldstein uses the estimates of P^* derived from his Cobb-Douglas production function, but estimates I^* indirectly using the definition $C^* = [P^*I^*]^{-1}$ and the estimates of C^* and P^*. He finds that the correlation between P^* and I^* is very weak ($r = 0.0048$) and that the sample variation in C^* reflects primarily the variation in P^*. Feldstein's 'costliness' index suffers from the same shortcomings as his index of technical efficiency, namely that it enables only a ranking of hospitals by costliness (it provides no information on how for hospitals' costs exceed their minimum feasible) and inevitably confounds inefficiency with random influences outside the hospital's control and 'statistical noise'.

2.4 NHS FACTOR INPUT MARKETS

The NHS is one of the largest employers in Western Europe: in 1983 it employed over 1m people (Annual Abstract of Statistics, 1986). The NHS is also a major purchaser of medical and other capital equipment, as well as pharmaceuticals, dressings and other 'disposables'. Despite its size, however, the NHS factor input market has been the subject of comparatively little econometric work. The issues of interest can be divided into demand-side issues and supply-side issues. Examples of the former include: to what extent do movements in relative factor prices cause changes in the demand for factor inputs? How far is the demand for factor inputs responsive to changes in 'desired' output levels? To what extent do NHS capital allocation decisions reflect politicians' attempts to 'buy' votes? Examples of supply-side issues include: how responsive is the supply of medical manpower to changes in relative wages? What variables apart from pay influence the supply of medical manpower and how?

Table 4 lists the principal British studies in this area to date.[10] There is inevitably a degree of arbitrariness in deciding what to include and what to exclude in a survey of the NHS factor input market. Table 4 excludes studies of the behaviour of suppliers of factor inputs other than manpower - it excludes, therefore, the pharmaceutical industry and the medical equipment industry - but includes studies of the NHS's demand for all factor inputs and of the supply of manpower to the NHS.

The paper by Gray et al. (1986) and the first two models of Lindsay (1980) are the principal studies of relevance to the demand side. The study of Gray et al. has already been discussed in section 2.3; it is included again here because it provides evidence on the way factor price movements have resulted in changes in the demand for factor inputs. Lindsay's first model is a conventional partial adjustment model of investment behaviour of the type encountered in studies of industry. His second model is more adventurous and is based on the hypothesis that politicians in power will use NHS funds to try to 'buy' votes in constituencies they expect to be marginal in the next general election. The two papers by Hoskins (1982a, b) and the third model of Lindsay (1980) are the principal studies of the supply of medical manpower to the NHS. The first of Hoskin's models is a fairly conventional aggregate-level model of labour supply, but the second is more sophisticated, involving the use of 'survivor' analysis of specific cohorts of manpower.

The results of the studies are discussed under two headings: (i) demand-side issues; and (ii) supply-side issues.

2.4.1 Demand-side issues

The study of Gray et al. and the first of Lindsay's two models both suggest that there is some substitution away from labour towards capital as the wage rate rises relative to the price of capital. The results of Lindsay suggest that a rise in the relative price of medical manpower of £1 per doctor p.a. produces an increase in desired capital expenditure of £4,120 (1970 prices). The results of Gray et al. also suggest that medical staff and professional/technical staff are fairly close complements, so that a rise in the price of one will tend to reduce the demand for other (at a given output level). Quite how reliable these results are is debatable. The weaknesses of the Gray et al. study were discussed in section 2.3. The study by Lindsay suffers from a number of shortcomings, some of which are indicated in Table 4. In this context it is worth noting that the relative factor price variable used by Lindsay refers only to physicians' and dentists' pay, and that the series is constructed on the assumption that the price of capital goods remained unchanged over the sample period of 28 years.

Only one study to date has attempted to relate factor input demand to changes in 'desired' output levels, namely the first of Lindsay's models. Though the capital expenditure series appears to relate to the NHS as a whole, the output measures used (like the majority of the other independent variables) refer only to the hospital sector. Actual output is a (1x3) vector comprising the number of inpatient cases, the number of outpatient cases and the number of accident visits. Desired output is related to actual output in three different ways: in the first actual and desired outputs are assumed to be equal; in the second desired output is a linear function of actual output; in the third desired output is a linear function of a three-period moving average of actual output. No adjustment for changes in casemix over time is made. The model selected is the model using the moving average-based output measures. Of the three coefficients on the output variables only the coefficient on the inpatient days variable was positive as predicted and implied that, for each additional inpatient year, capital expenditures would eventually increase by £266 (1970 prices).

One interesting issue is whether MHS funds are used by politicians to buy votes. This is the issue explored by Lindsay in his second model. The basic hypothesis of Lindsay's model is that the government will use NHS

Table 4: Studies of the NHS factor input market

Author(s)	Sample(s)	Details of model specification	Estimation method(s)	Remarks
1. Lindsay (1980) - Model I	Time-series data on NHS.	Partial adjustment model of investment, with expenditure on new capital as linear function of contemparaneous changes in desired 'output' and price of capital relative to price of medical manpower (in practice physicians only). 'Output' specified as vector of three output variables, based on inpatient cases, outpatient cases and accident visits.	OLS on data in levels and first-difference form; results based on data in first-difference form retained.	Data not well described: sample period not stated; sources vague; capital expenditure series seems to relate to NHS as a whole, but series for independent variables relate mainly to hospital sector.
2. Lindsay (1980) - Model II	Pooling of three cross-sections of data at Parliamentary constituency level for the general election years 1964, 1966 and 1970, and the years prior to each of these. For second equation (see right) data at Regional Health Authority also used.	Public choice model of investment based on hypothesis that government will use NHS funds to try to swing voting in marginal constituencies in its favour. Two-equation recursive system estimated: first equation explains how government determines which seats are likely to marginal; second relates NHS capital expenditures to inter alia expected closeness of result in local constituency, with 'expected closeness' based on results of first equation.	Both equations estimated by OLS on pooled data; since 'closeness' is a predicted variable, it is in effect an instrument, so that the second equation can be viewed as part of a simultaneous equation system with closeness and expenditure influencing one another.	
3. Lindsay (1960) - Model III	Time series data for UK, 1965-75.	Net emigration of physicians from UK related to current and lagged values of rate of return to employment in medicine. (Latter based on difference between GP's earnings and manual worker's earnings.)	OLS	
4. Hoskins (1982a)	Time-series data for England and Wales; 1953-75.	Model of supply of nursing manpower to NHS; derived from structural equations determining numbers joining and leaving in each year. Equation estimated is reduced-form equation, relating number of nurses at time t to number employed at time t-1, change in nurses' relative pay since time t-1, change in female unemployment rate, and female participation rate.	OLS; separate equations estimated for full-time State Registered Nurses (SRNs), full-time State Enrolled Nurses (SENs), part-time SRNs and part-time SENs.	
5. Hoskins (1982b)	Time-series data for England and Wales; 1950-69.	Analysis of supply of midwives to NHS using 'survivor' functions for specific cohorts of workers. Advantage over approach adopted by Hoskins (1982a) is that 'survivor' analysis can take into account that number of leavers will reflect changing cohort composition of total supply even in the absence of any changes in total supply.		
6. Gray et al. (1986)		See table 2.3		

funds to try to swing voters in marginal constituencies; the more marginal a constituency is expected to be, therefore, the greater will be the level of NHS capital expenditure in the constituency, ceteris paribus. The results of Lindsay's models - estimated on data at both Regional Health Authority level and Parliamentary constituency level - seem to provide some support for this hypothesis.

2.4.2 Supply-side issues

The results of the two studies by Hoskins (1982a, b) are not clear-cut on the question of pay. In the first study (confined to nurses) relative pay appeared to have an influence on supply only in the case of full-time nurses; in the case of part-time nurses supply seems to be relatively insensitive to (marginal) changes in pay. Quite how large the wage elasticity is in the case of full-time nurses is not clear. Indeed, in some specifications the elasticity was not significantly different from zero. The story was different in Hoskin's second study (concerning the supply of midwives): here relative pay turned out to be a key factor determining the proportion of a cohort surviving. A similar result was obtained by Lindsay in the third of his models listed in Table 4: he found that the (lagged) rate of return to employment in medicine was negatively (and significantly) related to net emigration by physicians from the UK.

The influence of factors other than pay is also somewhat unclear. Hoskins' first study suggests that the female unemployment rate (used as a proxy for labour market conditions) may be an important variable in determining the supply of part-time nursing staff, but appears to exert relatively little influence on the supply decisions of full-time staff. The latter result is consistent with the findings of Hoskins' second paper.

3. THE DEMAND FOR HEALTH CARE

On the supply side it is primarily the institutional details of the NHS that determines what issues are likely to be of interest from a policy perspective and what problems are likely to be encountered in exploring them. On the demand side institutional details play a much smaller role in this respect. More important are the (largely demand-side) peculiarities of the health care 'market'. As in the previous section, this section begins with a brief survey of the theoretical literature on the demand side: much of this is not of British origin, but much of it is as relevant in the NHS context as it is elsewhere. The following two sections introduce and summarize the results of the studies to date.

3.1 THEORETICAL CONSIDERATIONS ON THE DEMAND SIDE

The commodity 'health care' differs from other commodities in a number of key respects and these ought ideally to be taken into account in any analysis of the demand for health care. First, the demand for health care stems from a demand for a more fundamental commodity, namely health itself; the demand for health care is therefore a derived demand. This means that the demand for health care cannot really be analysed without simultaneously analysing the demand for those other inputs in the 'health production function' which are - to varying degrees - substitutable for health care.[11] In terms of figure 1 one cannot make much headway in investigating box C without at the same time considering box A.

The second complication is the so-called 'agency relationship'. In the words of Feldstein (1974): "Because the patient lacks the technical knowledge to make the necessary decisions, he delegates this authority to hs physician with the hope that the physician will act for him as he would for himself if he had the appropriate expertise" (Feldstein, op cit., p. 382). The physician thus in effect becomes both supplier and demander. The implications of this depend on whether the agency relationship is complete or incomplete. If it is complete (i.e., the physician acts solely in the interests of his patient), the agency model would be virtually indistinguishable from the textbook model of consumer behaviour. It is more likely, however, that the agency relationship is incomplete : physicians' decisions will typically reflect not only the preferences or their patients, but also their own self-interest, the pressures from professional colleagues, a sense of medical ethics and (possibly) a desire to make good use of available resources (cf, Feldstein, op cit.). One implication of an incomplete agency relationship is that health care utilization may well depend on the availability of facilities : when, for example, hospital beds are scarce, peer group pressure and a concern for the general welfare of patients in the community may induce physicians to restrict admissions and durations of stay to below what they would be if physicians were acting as 'perfect' agents. Resource availability may therefore be one of the key variables in explaining health care utilisation or 'demand'.

The third complication concerns the nature of the 'price' paid for health care. In most health systems the money price paid at the point of usage is often zero (or close to zero). (In the NHS user charges do exist in some areas of primary care - e.g., dental care and prescription drugs - but not for hospital care.) However, although there is frequently no money cost at the point of usage, there are other costs associated with the consumption of health care : in the UK these include income tax payments (part of which go towards the funding of the NHS), the time costs associated with treatment and any pecuniary losses suffered as a result of undergoing a course of medical treatment.

3.2 EMPIRICAL STUDIES OF THE DEMAND FOR HEALTH CARE

The principal issue that has dominated the British literature on the demand for health care is the availability effect. Is there a direct effect of availability on utilization and, if so, what is the elasticity of utilization with respect to availability? Partly due to the limited role money prices play in the NHS, very little research has been undertaken aiming to estimate price elasticities of demand. Neither has much research been done on the other determinants of demand, such as time prices, income and education. The studies that have been undertaken are listed in Table 5; again, the emphasis has been on the hospital sector, with only two of the studies listed being directed at the primary care sector.[12]

The results of the studies are discussed under four headings: (i) price elasticity of demand; (ii) income elasticity or demand; (iii) effects of availability; and (iv) other determinants of demand.

3.2.1 Price elasticity of demand

Money prices in the primary care sector (notably for dental care and prescription drugs) may effect utilization in three ways. First, they may deter patients from making the initial contact with their physician (or

Table 5: Studies of the demand for health care in the NHS

Author(s)	Sector	Sample	Demand measure(s)	Details of model specifications(s)	Remarks
1. Feldstein (1967) - Model I	Hospital Sector	NHS regions (excludes 4 London regions); England and Wales, 1960; n = 11.	No. admission; no. admissions + increase in waiting list; no. beds used; mean length of stay.	Demand related to no. available beds via linear, quadratic, exponential and double-log functions. Second set of equations estimated in which several additional independent variables were entered one at a time.	Argued that supply unlikely to be influenced by demand, because (i) most of the supply was created decades before 1960, (ii) NHS regions were natural catchment regions and (iii) no. persons living in region likely to be independent of available beds.
2. Feldstein (1976) - Model II	Maternity Care	Oxford Record Linkage Study of all deliveries in Oxford in 1962; n = 3491.	Admission to hospital (binary variable) and length of stay.	Two sets of models estimated. In first probability of admission to hospital modelled as function of (i) medical condition (age, parity, etc), (ii) socioeconomic group, (iii) marital status and (iv) availability of services. In second set of models length of stay related to same variables.	Details of estimated method used in admissions equations in Feldstein (1966).
3. Olowokure (1978)	Hospital Sector	Large acute non-teaching hospitals; England and Wales, 1967/68; n = 189.	Length of stay.	Mean length of stay related initially to 13 casemix variables and 16 'environmental' variables. In final specification only 7 casemix variables and 5 environmental variables included; latter were (in decreasing order of 'importance'): rate of inpatient operation, bed-population ratio, unit cost of medical pay, local authority welfare expenditure and rate of outpatient attendance.	
4. Frost and Francis (1979)	Hospital Sector	District Health Authorities within Trent Regional Health Authority; 1975; n = 17.	No. admissions.	Admissions related to available beds, no. consultants and population via double-log function.	Determinants of waiting lists also explored; see section 5.
5. Cullis et al. (1980)	Hospital Sector	2 sets of cross-section data: (i) as for Frost & Francis (1979): (ii) English Regional Health Authorities, 1979.	As Feldstein (1967).	As Feldstein (1967).	
6. Lavers (1983)	GP Care & prescription drugs.	Monthly time-series for UK; 1967-74.	Certified morbidity as proxy for GP consultations; volume of prescriptions dispensed.	Simultaneous equation model in which no. prescriptions , certified morbidity and average ingredient cost of prescription are determined simultaneously.	
7. Parkin and Yule (1985)	Dental Care	Time-series data for Scotland; 1962-81.	Patient-initiated contacts measured by no. examinations carried out and no. 'time-barred' estimates. Total utilization measured by: Gray's (1982) 'output' measure (cf, table 3.2); total no. procedures carried out; no. denture treatments; and no.	Utilization related to current and lagged prices, per capita income, dentists per 100,000 population, a time trend and various dummy variables (eg, for years when prices were frozen). Two alternative price variables employed: (i) maximum charge per course of treatment; and (ii) average charge paid by non-exempt patients.	

dentist). Second, they may influence physicians' prescribing behaviour once the contact has been made. Finally, they may deter patients from following the prescribed course of treatment. (In the case of prescriptions, for example, they may decide not to purchase the prescribed drug from the pharmacist.)

Evidence on the strength of these effects is mixed. The study by Lavers (1983) suggests that prescription charges do not have any deterrant effect on patients' consultation behaviour (as measured by certified morbidity), though the study by Parkin and Yule (1985) suggests that dental charges may have a (weak) deterrant effect on consultation behaviour. (In part, this difference may be due to the inclusion of current and lagged price variables in the Parkin-Yule equations.) On the question of total effects of prices on the volume of care these two studies are unequivocal: the elasticity of overall utilization (in volume terms) with respect to price is negative and significantly different from zero. This suggests that price may have a deterrant effect discouraging initial contacts, but it certainly has an effect once the patient has made the initial contact. The results do not provide any evidence to support the veiw that primary care physicians in the NHS respond to reductions in initial contacts by increasing the volume of care provided. The findings of Lavers suggest, however, that GPs respond to increases in prescription charges by prescribing either a greater quantity per prescription or more expensive items.

3.2.2 Income elasticity of demand

Only the two studies of Lavers (op cit.) and Parkin and Yule (op cit.) provide any evidence on the income elasticity of demand for health care in the NHS. Lavers found the income elasticity of demand for prescriptions to be positive and fairly large (0.79). Parkin and Yule, on the other hand, reported negative income elasticities in most of their specifications, implying literally that dental care is an inferior good; they suggest, though, that this may be spurious.

3.2.3 Effect of availability

As indicated in section 3.1, if the 'agency relationship' is imcomplete, the availability of health care facilities may have a direct effect on the consumption of health care. This possibility was first explored in the context of the NHS by Feldstein (1967), who found that a substantial proportion of the inter-regional variation in observed demand can be explained by availability differences. He also found that a linear equation relating demand to availability was more consistent with the data than either a quadratic or an exponential: demand appears, therefore, to increase with available supply but less than proportionately. The implication of this finding is that there is no level of supply at which the demand for beds would be satisfied. Broadly similar results were reported by Cullis et al. (1980), who replicated Feldstein's analysis using more recent and less aggregated data, and Frost and Francis (1979), who estimated a double-log function and found that the elasticity of hospital admissions with respect to available beds was not significantly different from unity. The importance of availability also comes out in Feldstein's (1967) individual-level analysis of maternity care and Parkin and Yule's (op cit.) study of the demand for dental care.

Whilst hospital admissions appear to be quite sensitive to bed availability, mean length of stay is, by contrast, relatively insenstive. Feldstein found that the elasticity of admissions with respect to bed avilability was roughly double the elasticity of mean length of stay (0.58 compared to 0.36).

3.2.4 Other determinants of demand

The studies in Table 5 shed relatively little light on the effects of other determinants of demand. There has been no attempt to date to investigate the role of education in the demand process, though Feldstein's (1967) finding that pregnant women from socioeconomic groups I and II have a higher probability of being admitted to hospital than women from lower socioeconomic groups may well reflect the role of education. Only one study to date has looked at the question of time prices (Parkin, 1980), but its conclusions are somewhat inconclusive. The role of age has also been only briefly explored: Feldstein (op cit.) found that age is positively associated with length of stay in hospital.

4. NON-MEDICAL INFLUENCES ON HEALTH

Over the course of the last few years it has become increasingly recognized that medical care is but one of the factors contributing towards good health. This has led to a growth of interest in the research going on in box A in Figure 1. Econometric research in this area can be divided into three interconnecting pigeon-holes: (i) 'health production functions'; (ii) the 'demand for health'; and (iii) studies investigating the demand for health-affecting commodities (e.g., cigarettes, food, exercise, etc) using approaches other than the 'demand-for-health' approach. The present survey is restricted to the literature in (i) and (ii).[13]

4.1 HEALTH PRODUCTION FUNCTIONS

Following the seminal work of Fuchs (1966), a strong tradition in health production functions has developed in the American literature. These studies view health as the outcome of a production process involving 'health inputs' such as medical care, behavioural variables, environmental factors and education; their results suggest that the marginal product of medical care in generating health improvements in the US is close to zero, but that the marginal products of other variables (education, in particular) are generally signficantly different from zero.[14]

The British work to date on health production functions has either been based on international data (see e.g., Cochrane et al., 1978) or has been focused on the effects of a limited set of independent variables. Feldstein and Butler (1965) and Forbes and Pickering (1985), for example, have investigated the effects of age and social class on perinatal mortality. A far more popular topic, however, has been the effects of unemployment on health: the British literature includes a large number of econometric studies of the unemployment-health link, as well as several papers discussing the methodological problems encountered in investigating the relationship.[15]

4.2 THE DEMAND FOR HEALTH

An extensive literature now exists on the demand for health, some of which originates from Britain.[16] This literature views the individual as possessing a stock of health capital that is subject to depreciation but which is capable of being augmented by acts of investment (health care, nutritious food, etc). None of the empirical literature in this area, however, has been based on British data.

5. MARKET EQUILIBRIUM AND NON-PRICE RATIONING

The absence of money prices in most parts of the NHS raises an obvious question, namely: if resources are not rationed according to willingness and ability to pay a market-clearing price, how are they rationed? This in turn prompts another question, namely: are resources rationed in such a way that the NHS tends towards an equilibrium? The answers to these questions are not obvious. It might be thought, for example, that in the absence of money prices excess demand would tend to grow forever by the amount of excess demand in each period. In practice, however, as Lindsay and Feigenbaum (1984) have noted, this does not seem to be the case, at least in the hospital sector: waiting lists for hospital care have grown very little over time. This suggests that the rationing device used may tend to move the system towards an equilibrium. How, though, are resources rationed? The answer to this question is clearly important in order to be able to predict the effects of different policy measures. It also has obvious equity implications.

5.1 THEORETICAL CONSIDERATIONS CONCERNING MARKET EQUILIBRIUM

Theoretical work on non-price rationing in the NHS may be broadly divided into (i) provider-orientated theories and (ii) consumer-orientated theories. It is convenient to begin with the former.

Frost (1977) has proposed a theory of non-price rationing for the hospital sector that emphasizes the role of the consultant in determining which patients are admitted for inpatient care. In Frost's model the key to the consultant's decision whether or not to admit is his view about the patient's state of health. He is assumed to have prior knowledge about the range of diseases that can give rise to the symptoms suffered by the patient and to be able to estimate the probability that the patient has each of the possible conditions. He is also assumed to have access to the results of diagnostic tests, on the basis of which he can arrive at 'posterior' probabilities. These will generally differ from his initial or 'prior' probabilities. The consultant decides whether to admit the patient or to refer him for outpatient treatment by comparing his own expected utilities associated with the two courses of action; he derives utility only in the event that his decision turns out to be correct and computes his expected utility using the posterior rather than the prior probabilities. If the expected utility of admission is greater than the expected utility of referral, he admits the patient.

The main prediction of this model is the (somewhat unsurprising) result that the higher the consultant's prior probability that the patient has a condition requiring inpatient treatment, the greater will be the likelihood that he will admit the patient. Frost suggests that one way a revision of the prior probabilities may come about is through a reduction in available capacity. The consultant can respond to a reduction in

capacity either by diagnosing at the same rate as before (and so accumulating a waiting list) or by revising upwards the threshold above which he operates (thereby leaving his waiting list unchanged); if he chooses the latter, fewer patients will be admitted. Frost inclines toward this latter view, suggesting that the consultant will adjust his prior probabilities to maintain his equilibrium. If this is true, a reduction in capacity will result in a reduction in the number of admissions and will leave the size of the waiting list unchanged. Frost also suggests that this implies that an increase in the number of consultants will result in an increase in the number of admissions and an increase in the length of waiting list (cf, Frost and Francis, 1979; Frost, 1980).

The rationing device in Frost's model is the consultant's set of prior probabilities, since it is these that are assumed to change to restore the consultant's equilibrium (cf, Frost, 1977, p. 798). This assumption is clearly rather unsatisfactory: it is stretching the imagination somewhat to argue that a physician responds to a reduction in hospital capacity by convincing himself that it is now less likely that the patient in front of him has a condition requiring inpatient treatment. It is clear too, though, that consultants may well alter their criteria for admission in response to capacity changes. Another - and altogether more plausible - story is that when physicians ration they do so not on the basis of 'prior probabilities' (which, if they are to mean anything at all, must remain constant in the face of capacity changes), but on the basis of some view of 'need'.

The role that 'need' plays as a rationing device on the NHS has been examined by Cooper (1974), who notes that it was the explicit aim of the founders of the NHS that health care should be rationed on the basis of 'need' rather than willingness and ability to pay a market-clearing price. The twin tasks of defining 'need' and discriminating between patients on the basis of need have been entrusted to the medical profession. What was not envisaged by the founders of the NHS is that 'need' is a relative rather than absolute concept, being one of many possible points along a continuum. In practice, the medical profession has tended to reassess its conception of need in line with actual levels of provision: an increase in throughput capacity results in physicians realigning their conception of need further along the continuum. More problematic, though, from the point of view of trying to develop a behavioural theory of resource allocation is Cooper's observation that different members of the medical profession have quite different concepts of 'need'. This, coupled with the clinical freedom the medical profession enjoys, means that rationing tends to take place in an uncoordinated and somewhat haphazard manner. The scope for generating testable predictions about resource rationing based on need is clearly, therefore, somewhat limited.

Theories of non-price rationing built up around the consumer have tended to focus on the role of time. One possibility is that time-prices may serve as a rationing device in the NHS (see Culyer and Cullis, 1976). It might be argued that, though NHS patients typically face no money price for the treatment they receive, they do face a time-price because they invariably have to wait for their treatment. An individual contemplating joining a waiting list for hospital treatment might therefore be assumed to do so only if the expected benefit of inpatient care exceeded the time-price. An implication of this is that if hospital throughput capacity increases, expected waiting time will fall, causing the time-price to fall and the number of people joining the waiting list to rise. Though Culyer and Cullis suggest otherwise, the size of waiting list need not, however, necessarily increase. It will tend to grow to the extent that the

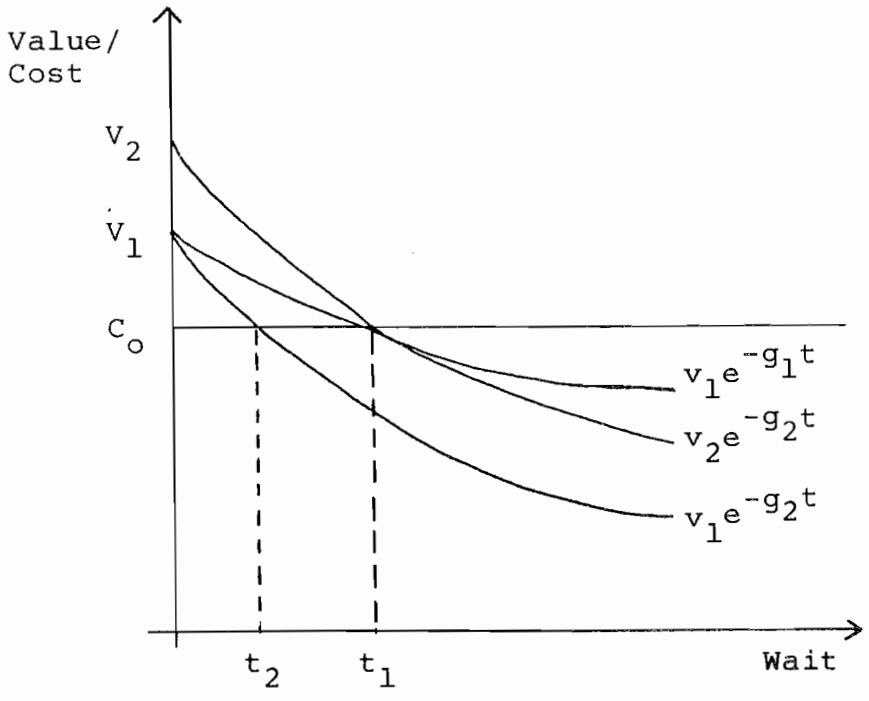
reduction in waiting time attracts new people onto the waiting list. This effect will be offset, however, by the shorter waiting time for those already on the list. If the demand for inpatient care is relatively inelastic with respect to waiting time, the former effect will be more than offset by the latter and the waiting list will fall.

The problem with the time-price argument is, as Culyer and Cullis note, that waiting for inpatient care does not usually impose any cost in the form of wasted time. Being on a waiting list does not normally prevent one from undertaking one's normal activities. In the context of NHS waiting lists, therefore, there is no in-kind time-price to act as a rationing device. This is not to say, though, that waiting does not impose any costs on the patient: he may suffer pain, discomfort and inconvenience during the wait; he may lose earnings; he, or his family, may be forced to bear additional costs (both pecuniary and non-pecuniary); and he may be subject to uncertainty. Waiting may, therefore, impose a time-price, but it is not an in-kind time-price. Providing this time-price is avoidable by not joining the waiting list, it may still act as a rationing device. In other words, providing patients have access to an alternative to NHS inpatient treatment (e.g., outpatient treatment, the private sector), the market demand for inpatient care will be sensitive to changes in expected waiting time.

Lindsay (1980) and Lindsay and Feigenbaum (1984) suggest an alternative theory of non-price rationing that also emphasizes the role of waiting time. In this model, however, waiting for inpatient care does not involve the payment of any time-price; instead, waiting makes inpatient care less attractive by reducing its value to the consumer. In part this is due to the effects of time preference. There is, however, another reason why waiting may reduce the present value of inpatient treatment, namely that the timing of the receipt of treatment is in itself important in determining the value of the treatment to the patient. For conditions which do not deteriorate over time, but which are nonetheless unpleasant, treatment today is clearly worth more (viewed from today) than treatment in six months time. It is for this reason that the demand for inpatient care will be sensitive to expected waiting time.

The position as viewed from the point of view of the consumer is illustrated in Figure 2.

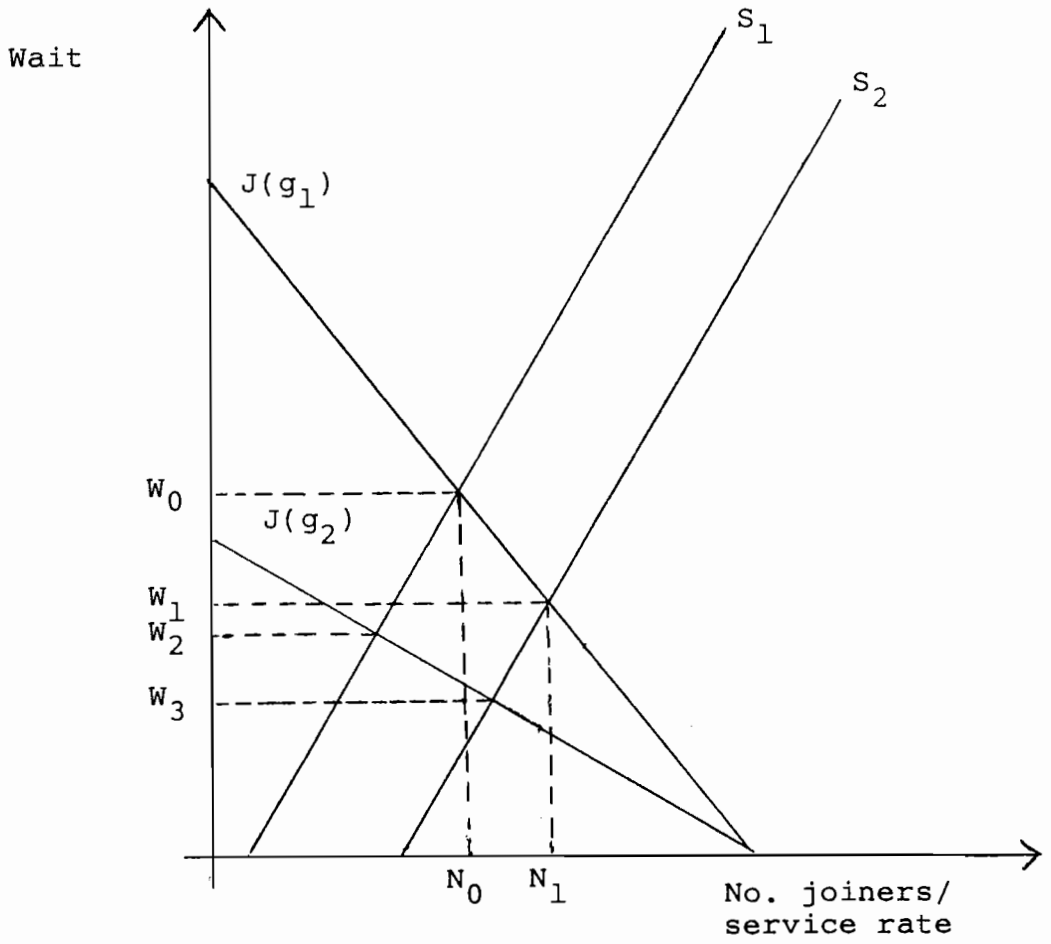
Figure 2



The downward-sloping curve $v_1 e^{-g_1 t}$ indicates the present value of inpatient treatment valued at v_1 for different expected waits; for the sake of simplicity the value is assumed to 'decay' at a constant rate g_1. In order to join the waiting list the individual has to incur a cost equal to c_o. If the value curve facing the individual is $v_1 e^{-g_1 t}$, he will join the waiting list only if the expected wait is not in excess of t_1. The marginal joiner is the consumer whose expected wait equals t_1 and whose costs of joining are exactly equal to the present value of the treatment. Changes in the 'decay' rate alter the slope of the 'v' function, but not its intercept on the vertical axis; the curve $v_1 e^{-g_2 t}$, for example, is associated with a higher decay rate than the curve $v_1 e^{-g_1 t}$. An increase in the decay rate will therefore reduce the critical wait from t_1 to t_2 for the consumer who values the treatment at v_1. The individual who was previously the marginal joiner will now, therefore, no longer join. The new marginal joiner will be the individual who places a value of v_2 on the treatment and whose 'v' curve is given by $v_2 e^{-g_2 t}$. An increase in the expected wait from t_2 to t_1 with a decay rate of g_2 means that the individual with the 'v' curve given by $v_1 e^{-g_2 t}$ who was previously the marginal joiner now becomes intra-marginal; the new marginal joiner is the consumer who values the treatment of v_2 and whose 'v' curve is given by $v_2 e^{-g_2 t}$.

On the basis of Figure 2 one can derive a joining function for the population representing the number of people joining the waiting list as a function of the expected waiting time, the average decay rate and the average value placed on the treatment. Figure 3 illustrates two joining functions, J_1 and J_2 for two different decay rates; in each case the number of joiners is inversely related to expected wait, but the number of joiners at any given wait is smaller the higher the decay rate.

Figure 3



The expected wait is determined not only by the number of joiners but also by the throughput (or 'service') rate of the hospital sector. In Figure 3 the two upward-sloping curves are supply or throughput curves; they are drawn on the assumption that if waiting time influences the throughput rate, its effect is positive. The equilibrium wait is determined by the intersection of the joining and throughput curves; for the two curves J_1 and S_1, for example, the equilibrium wait is w_o. At the equilibrium wait the number joining is equal to the throughput rate and the waiting list is numerically stationary. If the number joining is less than the throughput rate, the waiting list will shorten, waiting time will fall and new joiners will be attracted to the waiting list until the number of joiners is equal to the throughput rate. Conversely, if the number joining is greater than the throughput rate, the waiting list will lengthen, waiting time will rise and potential joiners will be discouraged from joining until the number of joiners is equal to the throughput rate. Waiting time thus automatically falls or rises until the waiting list is stationary and the equilibrium waiting time is attained. Waiting time therefore functions in much the same way as a market-clearing price.

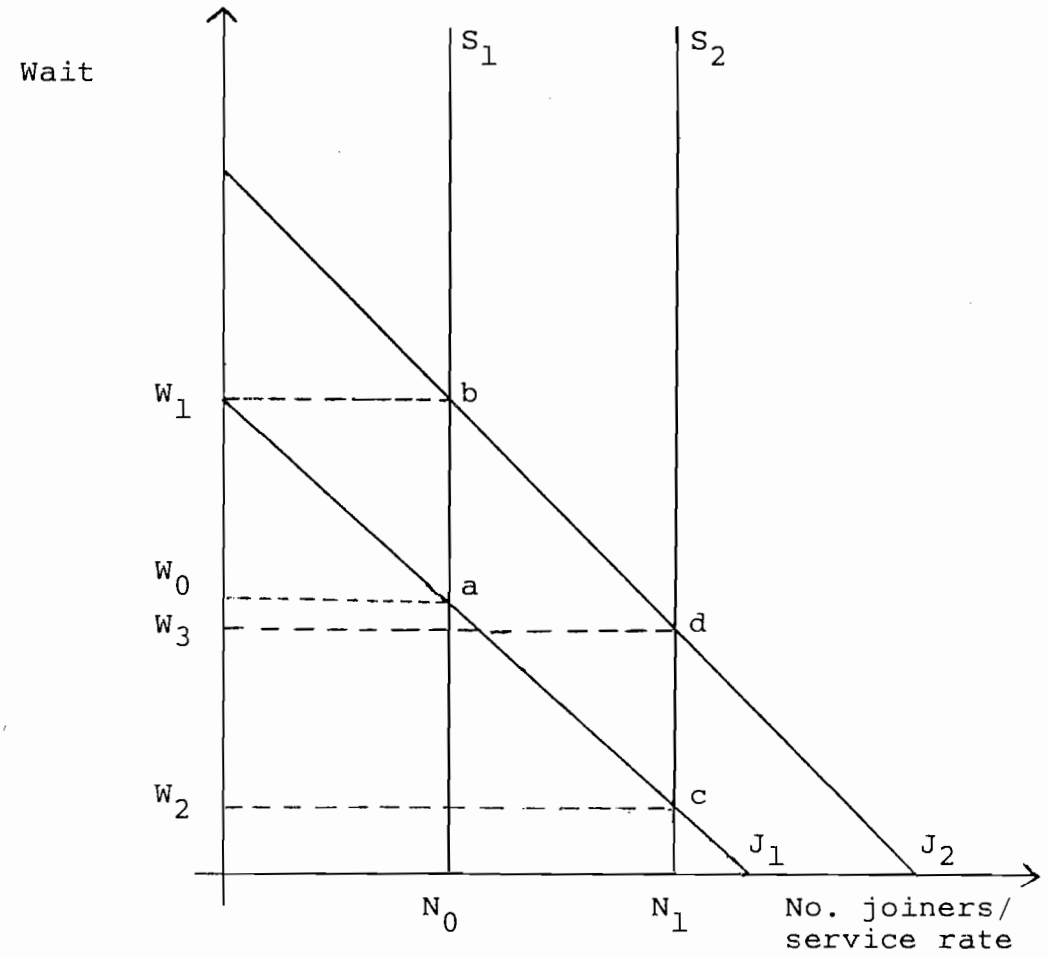
Using Figure 3 various comparative static results can be derived. An increase in throughput capacity (from, say, S_1 to S_2) reduces the equilibrium wait; the reduction in equilibrium wait is greater the lower the rate of decay (cf, the reduction from w_o to w_1 for the J_1 function and the reduction from w_2 to w_3 for the J_2 function). This reduction in waiting time results in a larger number of people joining the waiting list. Whether or not the size of the waiting list itself increases or falls depends on the elasticity of demand for inpatient care with respect to waiting time. This can be seen from Figure 3. The expected wait in any period is equal to the number on the waiting list divided by the throughput rate. In equilibrium the throughput rate is equal to the number of joiners, so that the equilibrium waiting list is equal to the product of the equilibrium waiting time and the equilibrium number of joiners; for the curves J_1 and S_1 in Figure 3, for example, the size of the waiting list is given by w_oJ_o. Whether w_1J_1 exceeds w_oJ_o depends on the elasticity of the J_1 function at w_oJ_o: if it exceeds one in absolute value, the waiting list will increase in size following the shift of the S curve from S_1 to S_2. An increase in the decay rate will cause a reduction in the number of joiners if supply is responsive to waiting time.

Before turning to the empirical work in this area it is worth considering the role of the physician in the consumer-orientated theories of non-price rationing. As Culyer and Cullis (op cit.) and Lindsay (op cit.) note, none of the analysis above is undermined by the fact that it is the physician who refers the patient to hospital and assigns him to a waiting list, providing the physician's judgement of the value of the treatment and of the costs involved is broadly the same as that of the patient. What, though, if the 'agency relationship' is incomplete and the physician exploits his position of superior knowledge to his advantage? Evans (1974), for example, has suggested that physicians might react to an increase in their numbers by shifting their patients' demand curves outwards, thereby offsetting any reduction in the demand for their services. If this is true in the context of the NHS, market clearing would take place directly through the information passed by physicians to their patients as well as by adjustments on waiting time.

Lindsay (op cit.) proposes a test to detect the existence of this 'supplier-induced' demand (SID). The initial situation in Figure 4 is at a point where equilibrium time is w_o. An increase in the number of physicians would, according to the SID theory, result in an outwards shift

in the joining function from J_1 to J_2, causing the equilibrium waiting time to rise to w_1. The 'conventional' view, Lindsay suggests, is that the increase in the stock of physicians would result in an outwards shift in the throughput curve from S_1 to S_2, causing equilibrium waiting time to fall to w_2. Lindsay's proposed test is therefore based on the estimated effect of a change in the stock of physicians on waiting time: a positive effect would support the SID theory, whilst a negative effect would support the 'conventional' view. This proposed test is not entirely satisfactory, since it is unreasonable to assume that throughput is unchanged under the SID hypothesis (cf, Cullis and Jones, 1985); if the J and S functions shift

Figure 4: Supplier-induced demand in Lindsay's model of non-price rationing



simultaneously, the new equilibrium waiting time could still be lower than the old even in the presence of SID (e.g., w_3 instead of w_o). Thus whilst an increase in waiting time is consistent only with the SID hypothesis, a reduction in waiting time is not incompatible with the existence of SID (cf, Maynard, 1983).

5.2 EMPIRICAL STUDIES OF MARKET EQUILIBRIUM

Before turning to the empirical evidence on non-price raitoning in the NHS it may be useful to summarize the predictions of the theories outlined in the previous section. The principal predictions are as follows: (i) the number of persons joining the waiting list in each period is predicted to be inversely related to waiting time by both the time-price (TP) and demand-decay (DD) theories; (ii) according to the DD theory the number joining at any given wait is smaller the higher the decay rate; (iii) both the DD and TP theories predict an inverse relationship between throughput capacity and waiting time; (iv) according to the DD theory the reduction in waiting time resulting from an increase in capacity will be greater the lower the decay rate; (v) the DD and TP theories both predict a positive (negative) effect of throughput capacity on waiting list if the elasticity of demand for admissions with respect to waiting time is greater (less) than unity in absolute value and no effect if the elasticity is exactly one. Frost's consultant-based theory predicts that there will be no relationship throughput capacity and waiting list; (vi) a positive relationship between the number of consultants and waiting time is consistent with the SID version of the DD theory, but not with the consumer sovereignty version. A negative relationship is consistent with both versions; (vii) the relationship between the number of consultants and waiting list can be negative or positive under both versions of the DD theory. Frost's theory predicts a positive relationship.

Table 6 lists the empirical studies to date which shed light on the issue of non-price rationing. Not all provide evidence on all of the predicitons listed above. Only the study of Lindsay and Feigenbaum (1984) provides any evidence on (i) above: the coefficient on waiting time in the joining function is negative as predicted, but is significantly different from zero only at the 90% level. The same is broadly true for prediction (ii): Lindsay and Feigenbaum reported a negative coefficient on the interaction of waiting time and the decay rate but the corresponding t-statistic is always less than 2.0 and generally less than 1.90. The evidence on prediction (iii) is rather more conclusive: both Culyer and Cullis (1976) and Lindsay (1980) found significant negative relationships between throughput capacity and waiting time. The results of Lindsay also lend support to prediction (iv): significantly larger capacity effects were found for low decay conditions. Turning to prediction (v), most of the studies to date that have estimated equations for waiting lists have reported negative but insignificant coefficients on bed availability; the exception is the equation for General Medicine in Cullis *et al*. (1980) where a positive and significant coefficient was reported. Lindsay and Feigenbaum estimate the elasticity of demand for inpatient care with respect to waiting time at between -0.70 and -0.55, but no tests were undertaken to determine whether the estimates are significantly different from -1.0. Nonetheless, the results of the waiting list studies and the estimates of the waiting time elasticity would appear to be broadly consistent with the TP and DD theories. They are, however, also consistent with Frost's theory. Only one study to date - that of Lindsay (*op cit*.) - provides evidence on prediction (vi), reporting a negative relationship between the number of physicians and waiting time: this result is Table 6

Table 6: Empirical studies of non-price rationing on NHS

Author(s)	Sample(s)	Details of model specification	Estimation method	Remarks
1. Feldstein (1967)	NHS regions (excludes 4 London regions); England and Wales, 1960; n = 11.	Waiting list related to beds via linear and quadratic equations.	OLS.	
2. Culyer and Cullis (1975)	Time-series and cross-section data based on NHS regions.	3 sets of equations estimated: (i) waiting list related to beds via linear and quadratic equations; (ii) waiting list related to beds, consultants, registrars and junior doctors; (iii) change in waiting list related to previous year's change in waiting list and beds.	OLS.	First set of equations replicate Feldstein's (1967) work.
3. Culyer and Cullis (1976)	As Culyer and Cullis (1975).	Correlation coefficients between (i) waiting time and beds; (ii) waiting list and beds; and (iii) waiting list and waiting time.		Test of time-price theory of non-price rationing.
4. Frost and Francis (1979)	17 districts within Trent Regional Health Authority; General Surgery only; 1975.	Waiting list related to: (i) consultants; (ii) consultants and beds; and (iii) consultants, beds and population. Double-log functions used throughout.	OLS; Restricted Least Squares.	Cross-section test of Frost's consultant-based theory of non-price rationing.
5. Cullis et al. (1980)	2 sets of cross-section section data: (i) as for Frost and Francis (1979); (ii) English Regional Health Authorities 1979.	Waiting list related to beds via linear equation.	OLS	Replication of Feldstein's (1967) work. Quadratic also estimated but 'failed' due to multicollinearity.
6. Frost (1980)	Time-series data for England and Wales, 1949-76; Most data refer to General Surgery.	Partial adjustment model in which equilibrium waiting list depends on no. consultants. In same versions of estimating equations other variables included; eg, effective hospital capacity as measured by available beds ÷ length of stay.	Maximum likelihood.	Time-series test of Frost's (1977) consultant-based model of non-price rationing.
7. Lindsay (1986)	14 Regional Health Authorities of NHS, 1974; observations based on ICD categories.	2 sets of equations estimated: (i) equations relating waiting time to beds and change in bed stock since previous year, with different equations estimated for each of the two 'decay' categories of disease; (ii) waiting time related to beds, physicians per 1000 population, dummy for disease category and beds x dummy.	OLS	Diseases divided into high decay, low decay and negative decay categories: high decay category includes non-emergency cases that are amenable to drug therapy where alternative to hospital treatment is available; low decay category includes cases that do not deteriorate but for which no alternative to hospital treatment is available; negative decay category includes emergency cases.
8. Lindsay and Feigenbaum (1984)	As Lindsay (1980)	Joining equation estimated by relating admissions to waiting time and dummies for age structure of population. Supply equation also estimated relating admissions to physicians per 1000 population, available beds and (predicted) waiting time.	IV; predicted values of waiting time used, derived from regressions of waiting time on bed availability, no. physicians per 1000 population & dummy variable for decay category.	Parameter estimates for different queues made comparable by normalizing on basis of ratio of estimated potential joiners in the two queues.

consistent with both the SID and consumer sovereignty versions of the DD theory. The evidence on prediction (vii) is mixed. Culyer and Cullis (1975) found no relationship between the number of consultants and the length of waiting list, but Frost and Francis (1979) and Frost (1980) reported positive relationships with the corresponding elasticities equal to 1.0. Only Frost's theory predicts that the relatinship will definitely be positive.

6. PLANNING, BUDGETING AND MONITORING MECHANISMS

Econometric studies that are designed specifically to assist in health care planning fall into two broad groups: the first comprises the literature on multi-equation models of the health care sector and the second includes the literature on linear programming. The literature in both groups dates back to Feldstein's (1967) early work on British data. The British studies in each group are discussed in turn.

6.1 ECONOMETIC MODELS OF THE HEALTH CARE SECTOR

The only sector-wide model of the NHS to date is Feldstein's (1967) aggregate planning model.[17] The model consists of a system of nine equations that together determine: (i) per capita levels of local authority expenditure on midwifery, health visiting , home nursing and domiciliary care; (ii) the availability of GP services; and (iii) the rate of admission to hospital and mean length of stay. Two of the model's equations are identities. The predetermined variables in the model are: the proportion on the population aged over 75; the ratable value of property; the proportion of females in the population; the proportion of the population in the lower socioeconomic groups; the proportion of children under 15 years of age; population density; GP availability lagged six years; local authority expenditure lagged six years; beds in the county; beds in the broader area; and mean per capita income. Variables are defined on a per year and (where appropriate) per capita basis and relate to the counties of England and Wales for 1960.

The estimates of the models structural parameters are reproduced in Table 7. The results for the two hospital sector equations (hospital admissions and mean stay) generally support the results obtained by Feldstein reported in section 3.2: hospital admissions are highly sensitive to local supply conditions, but mean stay is less sensitive. Also of interest in the context of hospital care are the findings regarding the effects of GP availability and local authority health expenditure: both tend to increase the number of hospital admissions, though neither coefficient is statistically significant. Both hospital admissions and mean stay are positively related to the proportion of elderly persons and females in the local population. Turning to the primary care sector, it is interesting to note that none of the coefficients on the hospital bed variable is significantly different from zero, suggesting that the degree of substitutability between hospital care and primary care may be limited. The coefficients on GP availability in the home nursing and domiciliary care equations on the other hand suggest a significant degree of complementarity between GP care and these two primary care services.

One attraction of a model such as this is that it enables one to predict the effects of changes in policy variables. As formulated, Feldstein's model contains only one policy instrument under the direct control of central government, namely the stock of beds in general

Table 7: Structural parameter estimates of Feldstein's (1967) health sector model

	MIDWIFERY	HOME NURSING	DOMICILLARY CARE	HEALTH VISITING	GP AVAILABILITY	HOSPITAL ADMISSIONS	MEAN STAY
ELDERLY (75+)		0.26	-0.21		-0.06	0.20	0.38
PROPERTY VALUE	-0.71*	-0.08	-0.04	-0.01			
FEMALES	4.76*			2.27		0.85	1.96
SOCIAL CLASS	0.73*	-0.26	-0.26	0.23	0.20		
CHILDREN	1.80*			0.68	-0.97*		
POPULATION DENSITY	0.06	-0.01	-0.00	-0.09	-0.03	0.04	-0.01
BEDS USED+	-0.48	-0.39	-0.01	0.23			
GP AVAILABILITY+	0.07	0.49*	0.29*	0.13		0.14	
GP AVAILABILITY$_{-6}$					0.67*		
LA EXPENDITURE$_{-6}$	0.24	0.60*	0.53*	0.64*			
BEDS AVAILABLE (COUNTY)					-0.06	0.20*	0.14
BEDS AVAILABLE (AREA)					0.48	0.26	0.15
MEAN STAY+						-0.08	
HOSPITAL ADMISSIONS+							0.26
LA HEALTH EXPENDITURE+					0.22	0.20	-0.03
PER CAPITA INCOME							-0.10

* Statistically significant at 95% level (two-tail test)

\+ Treated as endogenous variable.

hospitals. The model's reduced-form parameters relating to bed availability suggest that increases in the stock of beds result in reductions in local authority expenditure on midwifery and home nursing, but cause increases in expenditure on domiciliary care and health visiting. GP availability also rises with increases in bed supply. Hospital admissions and mean stay also increase with bed availability, but admissions increase by a greater proportion.

Though bed availability is the only policy instrument under the control of central government, Feldstein also estimates the reduced-form parameters for variants of the model in which first GP availability and then local authority health expenditure are treated as policy instruments. The results of the both are broadly similar and may be summarized as follows. Increases in bed availability cause a larger reduction in home nursing than before, but now also produce a decrease in expenditure on domiciliary care. Health visiting also rises less in response to increases in bed availability than before. These results derive from the fact that increased bed availability no longer attracts additional GPs who would have caused local authority expenditure to rise. In this specification of the model the admission and mean stay elasticities with respect to bed availability are almost identical in absolute magnitude.

6.2 APPLICATIONS OF LINEAR PROGRAMMING

The British health economics literature to date contains only one empirical study employing linear programming, namely Feldstein's (1967) model of casemix planning. The hospital's problem is to determine its optimal casemix given a limited budget and limited availability of certain inputs. Casetypes are defined according to department and outputs (cases treated) are related to inputs via fixed input-output coefficients; there are four inputs (nurses, doctors, bed-days and 'purchasables') and nine casetypes. Hospitals are assumed to have fixed quantities of the first three inputs and expenditure on purchasables is assumed to be constrained by the budget available and the unit costs of nursing and medical staff. The solution to the planning problem is found by maximizing the hospital's objective function - assumed to be a weighted function of the number of cases treated in each casemix category - subject to the technology and budget constraints.

The coefficients of the production technology are estimated via regression analysis using data for large, acute non-teaching hospitals in England and Wales for the year 1960/61. In specifying the objective function Feldstein suggests that the treatment of additional cases should be subject to diminishing marginal valuation, since some cases will benefit more from hospital treatment than others. To capture this the weights attached to the treatment of each casetype are specified as monotonically decreasing step functions with an infinite value being placed on the first n_{1j} cases of the jth casetype and a zero value on all cases after n_{2j}. This constrains hospitals to treat at least n_{1j} cases of type j but not more than n_{2j} cases. Three different sets of weights are used: the first set are all equal to one another; the second are proportional to average expenditure per casetype; and the third are proportional to average length of stay.

The optimal casemix is found to be highly sensitive to the choice of objective function. When all cases are given equal weight the results indicate that minimum numbers should be treated in three of the nine casemix categories and maximum numbers in three of the remaining six

categories. The limited availability of nurses, doctors and budget are all binding constraints, but slack exists in the case of bed-days. The shadow price of doctors is the highest of the shadow prices, implying that additional expenditure on doctors is more productive than expenditure on other inputs. Different results, however, emerge when different casetypes are not weighted equally. When lengths of stay are used as weights all output above the minimum is concentrated on medical cases and the bed-days constraint is the only binding constraint. However, when average costs are used as weights output above the minimum is concentrated on medical and surgery cases and both the budget constraint and bed-days constraint are binding. Feldstein emphasizes that, in view of the sensitivity of the results to the choice of objective function, the exercise ought to be seen only as an illustration of how linear programming might be used in casemix planning.

7. WHITHER NOW?

Scope exists for further applied econometric work in each of the boxes in Figure 1. The following list of suggestions for future research has been drawn up with an eye to the main themes in the current debate on British health policy.

On the supply side a major hindrance to effective applied econometric work is the dearth of theoretical literature on provider behaviour. At a time when there is so much interest in the effects of changes in the incentive structures facing health care providers, the lack of theoretical literature of relevance is particularly unfortunate.[18] There is a wide range of questions that merit attention here: how would clinicians be expected to respond to the introduction of clinical budgeting? How would hospitals respond to the introduction of a prospective reimbursement payment system? What effects would changes in the system of remunerating GPs have on consultation rates, referral patterns, casemixes and choice of location? How far does the current system of remunerating dentists encourage the provision of 'unnecessary' care? These are all questions that call for the construction and testing of models of provider behaviour; in some instances the data may need to be generated from large-scale experiments. Still on the supply side, there is scope for further work with production functions and cost functions, not only on the hospital sector but also on the primary care sector. With the current emphasis on efficiency in the NHS further research is badly needed on the measurement of efficiency, as well as on the issue of where improvements might come from. On the question of efficiency measurement the stochastic frontier model may well prove a useful tool: the stochastic cost frontier model, for example, would enable the more serious of the problems inherant in Feldstein's (1967) 'costliness' index to be overcome. In view of the current interest on the part of the DHSS in this index (see e.g., Smith, 1983), this would seem to be an avenue worth exploring.[19] On the question of where efficiency improvements might come from factor substitution (and particularly manpower substitution) ought probably to be a top priority; does the growing use of ancillary staff by GPs, for example, represent a move towards a more efficient factor mix and, if so, how far is there to go before an efficient factor mix is reached? The issues of economies of scale and short-run average versus marginal costs would also seem to merit some attention; in investigating the former the limitations of the studies to date and the suggestions of Cowing *et al.* (1983) should be borne in mind. Finally, more econometric research would seem called for on the NHS factor input market; empirical work on the determinants of the demand for and supply of factor inputs has really only just begun.

On the demand side much more research into the determinants of demand is needed. Far too little information is available on the deterrant effects of pricing; in the light of recent proposals to extend pricing in the NHS, this is to be regretted. Information on the other determinants of demand (particularly in primary care) is also far too limited. Here applied econometric work has a potentially major contribution to make to the current debate on socioeconomic inequalities in health care utilization.[20]

The almost complete absence of any British work to date on health production functions and the demand for health is a great pity. Health production functions would seem to offer a chance of advancing the current debate on the alleged under-funding of the NHS beyond the accounting-type exercises that have dominated the debate so far.[21] The demand-for-health approach, on the other hand, affords insights into a whole range of issues, including: socioeconomic inequalities in health;[22] the formulation of effective prevention policies;[23] the impact of health on labour supply, wage rates and retirement decisions;[24] and the health consequences of unemployment.[25]

The issue of market equilibrium and non-price rationing would also seem to warrant further research. Far too little of the research effort to date has been directed at exploring the testable implications of different theories of non-price rationing. In view of the different equity implications of different theories of non-price rationing and the importance of having a model of the rationing process in order to be able to explore the consequences of different public policies, the testing of rival theories ought probably to be a high priority for applied econometric work in this area.

The development of econometric models for health sector planning ought also to be a high priority for the future. Econometric models of the health care sector have been widely employed in the United States to simulate the effects of major changes in public policy (see e.g., Friedman and Pliska, 1985); a sector-wide model of the NHS might prove useful for forecasting the consequences of exogenous changes, such as population ageing, as well as for policy simulation. Linear programming might also be more profitably employed than it has been to date; Lavers's (1972) suggestion of using linear programming to elicit implicit valuations in current decisions in the health care sector would also seem to be worth following up.

There are, in short, many avenues that are still to be explored. In view of the sheer size of the NHS and the almost unique opportunities it affords for developing and testing econometric models in an environment where non-price rationing devices dominate, it would be a great pity if these avenues remained unexplored.

FOOTNOTES

1. For state-of-the-art surveys with a slight British bias see Williams (1977) and Culyer (1987).

2. For a guide to the English language literature on economic appraisal see Drummond (1981) and Drummond et al. (1986).

3. A survey of the American literature is to be found in Feldstein (1974).

4. The survey covers discussion papers and unpublished doctoral dissertations, as well as published material, but excludes unpublished papers, such as papers presented to the biannual meetings of the Health Economists Study Group. It also excludes econometric studies of the personal social services; a survey of these is to be found in Knapp (1984).

5. For surveys of models of hospital behaviour see Jacobs (1974) and McGuire (1985). On models of physician behaviour see Pauly (1980).

6. The best known American studies are those of Reinhardt (1972), Pauly (1980) and Jensen and Morrisey (1986a, b).

7. Unlike the Cobb-Douglas, the translog allows the Allen (partial) elasticity of substitution to vary between pairs of factors and does not constrain all factors to be substitutes (i.e., complementarity is permitted).

8. At the same time it needs to be acknowledged that estimating the translog directly invariably gives rise to severe multicollinearity and - as with OLS estimation of the Cobb-Douglas - probably gives rise to some simultaneous-equation bias.

9. Table 3 excludes several studies in the Operations Research literature (see e.g., Coverdale et al., 1980; Ashford et al., 1981; Bailey and Ashford, 1984). None of these, however, is based on the economic theory of the cost function, all being essentially cost equations rather than cost functions. For a recent survey of the North American literature on cost functions see Cowing et al. (1983).

10. Feldstein (1974) is probably the best source for American studies of health care factor input markets.

11. Grossman (1972) is the seminal reference on the derived nature of the demand for health care.

12. For surveys of the American literature on the demand for health care and dental care see Newhouse (1981) and Yule and Parkin (1985) respectively.

13. A survey of the English language literature on the demand for alcohol and cigarettes is to be found in Godfrey (1986).

14. Best known amongst these studies are those of Auster et al., (1969), Benham and Benham (1975), Newhouse and Fridlander (1980) and Hadley (1982).

15. Surveys of the empirical studies to date are to be found in Stern (1983) and Wagstaff (1985a). Recent studies include those of Forbes and McGregor (1984) and Narendranathan et al., (1985). Discussions of the methodological problems are to be found in Gravelle (1984), McAvinchey (1984) and Wagstaff (1986a).

16. For surveys of the literature on the demand for health see Grossman (1977, 1982) and Muurinen (1982). Recent contributions include Dardanoni and Wagstaff (1987), Wagstaff (1986b) and Wolfe (1986).

17. A survey of econometric models of the United States health care sector is to be found in Friedman and Pliska (1985).

18. See, for example, Maynard et al. (1986) and Wickens and Cole (1985).

19. For an application of the stochastic cost frontier model to British crematoria see Hammond (1986).

20. See LeGrand (1978, 1982) and Collins and Klein (1980).

21. See Bosanquet (1985).

22. On the controversy surrounding the measurement of inequalities in health see Townsend and Davidson (1980) and Illsley and LeGrand (1987). On trying to explain inequalities in health via the demand-for-health approach see Muurinen and LeGrand (1984) and LeGrand (1985).

23. On the inter-relationships between health, education and time preference see Grossman (1975) and Fuchs (1978, 1982).

24. See Grossman and Benham (1984) and Wolfe (1986).

25. On trying to model the health consequences of unemployment via the 'Grossman' model see Wagstaff (1985b).

REFERENCES

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977). Formulation and estimation of stochastic production function models, **Journal of Econometrics 6**, 21-37.

Ashford, J.R., M.S. Butts and T.C. Bailey (1981). Is there still a place for independent research into the issues of public policy in England and Wales in the 1980s? **Journal of the Operational Research Society 32**, 851-64.

Auster, R., I. Leveson and D. Sarachek (1969). The production of health: an exploratory study, **Journal of Human Resources 4**, 411-36.

Bailey, T.C. and J.R. Ashford (1984). Specialty costs in English hospitals: a statistical approach based on a cost component model, **Journal of the Operational Research Society 35**, 247-56.

Barlow, R. (1968). Review of 'Economic Analysis for Health Service Efficiency', **Economic Journal 78**, 921-3.

Benham, L. and A. Benham (1977). The impact of incremental medical services on health status 1963-70, in: R. Anderson et al., eds., **Equity in Health Services: Empirical Analysis in Social Policy**, (Ballinger, Cambridge Mass.).

Berndt, E.R. and L.R. Christiensen (1973). The translog function and the substitution of equipment, structures and labor in US manufacturing 1929-68, **Journal of Econometrics 1**, 81-114.

Bosanquet, N. (1985). **Public Expenditure on the NHS**, (IHSM, London).

Cochrane, A.L., A.S. St.Leger and F. Moore (1978). Health service 'input' and mortality 'output' in developed countries, **Journal of Epidemiology and Community Health 32**, 200-5.

Collins, E. and R. Klein (1980). Equity and the NHS: self-reported morbidity, access and primary care, **British Medical Journal 281**, 1111-5.

Cooper, M.H. (1974). Economics of need: the experience of the British National Health Service, in: M. Perlman, ed., **The Economics of Health and Medical Care**, (Macmillan, London).

Coverdale, I., R. Gibbs and K. Nurse (1980). A hospital cost model for policy analysis, **Journal of the Operational Research Society 31**, 801-11.

Cowing, T.G., A.G. Holtmann and S. Powers (1983). Hospital cost analysis: a survey and evaluation of recent studies, in: R.M. Scheffler and L.F. Rossiter, eds., **Advances in Health Economics and Health Services Research Volume 4**, (JAI Press, Connecticut).

Cullis, J.G., D.P. Forster and C.E.B. Frost (1980). The demand for inpatient treatment: some recent evidence, **Applied Economics 12**, 43-60.

Cullis, J.G. and P.R. Jones (1985). National Health Service waiting lists: a discussion of competing explanations and a policy proposal, **Journal of Health Economics 4,** 119-35.

Culyer, A.J. (1987). The future of health economics in the UK, in: G. Teeling-Smith, ed., **Health Economics: Prospects for the Future,** (Croom Helm, London).

Culyer, A.J. and J.G. Cullis (1975). Hospital waiting lists and the supply and demand of inpatient care, **Social and Economic Administration 9,** 13-25.

Culyer, A.J. and J.G. Cullis (1976). Some economics of hospital waiting lists in the NHS, **Journal of Social Policy 5,** 239-64.

Culyer, A.J. and M.F. Drummond (1978). Financing medical education: inter-relationships between medical school and teaching hospital expenditure, in: A.J. Culyer and K.G. Wright, eds., **Economic Aspects of Health Services,** (Martin Robertson, London).

Culyer, A.J., R.J. Lavers and A. Williams (1971). Social indicators: health, **Social Trends 2,** 21-41.

Culyer, A.J. and A.K. Maynard (1981). Cost-effectiveness of duodenal ulcer treatment, **Social Science and Medicine 15C,** 3-11.

Culyer, A.J., J. Wiseman, M.F. Drummond and P. West (1978). What accounts for the higher costs of teaching hospitals? **Social and Economic Administration 12,** 20-30.

Culyer, A.J., J. Wiseman, M.F. Drummond and P. West (1982). Revenue allocation by regression: a rejoinder, **Journal of the Royal Statistical Society Series A 145,** 127-33.

Dardanoni, V. and A. Wagstaff (1987). Uncertainty and the demand for medical care, Discussion Paper 28, (Centre for Health Economics, University of York).

Davis, K. (1968). Review of 'Economic Analysis for Health Service Efficiency', **American Economic Review 58,** 1488-90.

Drummond, M.F. (1981). **Studies in Economic Appraisal in Health Care,** (Oxford University Press, Oxford).

Drummond, M.F., A. Ludbrook, R. Lowson and R. Steele (1986). **Studies in Economic Appraisal in Health Care, Volume 2,** (Oxford University Press, Oxford).

Evans, R.G. (1971). Behavioural cost functions for hospitals, **Canadian Journal of Economics 4,** 198-215.

Evans, R.G. (1974). Supplier-induced demand: some empirical evidence and implications, in: M. Perlman, ed., **The Economics of Health and Medical Care,** (MacMillan, London).

Evans, R.G. and M.D. Walker (1972). Information theory and the analysis of hospital cost structure, **Canadian Journal of Economics 5,** 398-418.

Feldstein, M.S. (1965). Hospital cost variation and casemix differences, **Medical Care 3,** 95-103.

Feldstein, M.S. (1966). A binary variable multiple regression model of analysing factors affecting perinatal mortality and other outcomes of pregnancy, **Journal of the Royal Statistical Society Series A 129,** 61-73.

Feldstein, M.S. (1967). **Economic Analysis for Health Service Efficiency: Econometric Studies of the British National Health Service,** (North-Holland, Amsterdam).

Feldstein, M.S. (1974). Econometric studies of health economics, in: M.D. Intriligator and D.A. Kendrick, eds., **Frontiers of Quantitative Economics Volume 2,** (North-Holland, Amsterdam).

Feldstein, M.S. and N.R. Butler (1965). Analysis of factors affecting perinatal mortality: a multivariate statistical approach, **British Journal of Preventive and Social Medicine 19,** 128-34.

Forbes, J.F. and A. McGregor (1984). Unemployment and mortality in post-war Scotland, **Journal of Health Economics 3,** 234-57.

Forbes, J.F. and R. Pickering (1985). Influence of maternal age, parity and social class on perinatal mortality in Scotland 1960-82, **Journal of Biosocial Science 17,** 339-49.

Friedman, B. and S.R. Pliska (1985). Hospital expenses in a sector modal, **Health Services Research 19,** 717-52.

Frost, C.E.B. (1977). Clinical decision-making and the utilization of medical resources, **Social Science and Medicine 11,** 793-9.

Frost. C.E.B. (1980). How permanent are waiting lists? **Social Science and Medicine 14C,** 1-11.

Frost, C.E.B. and B.J. Francis (1979). Clinical decision-making: a study of General Surgery within Trent RHA, **Social Science and Medicine 13A,** 193-8.

Fuchs, V.R. (1966). The contribution of health services to the American economy, **Milbank Memorial Fund Quarterly 4,** 65-101.

Fuchs, V.R. (1969). Review of 'Economic Analysis for Health Service Efficiency', **Health Services Research 3,** 242-50.

Fuchs, V.R. (1979). The economics of health in a post-industrial society, **The Public Interest 56,** 3-20.

Fuchs, V.R. (1982). Time prefernce and health: an exploratory study, in: V.R. Fuchs, ed., **Economic Aspects of Health,** (NBER, New York).

Godfrey, C. (1986). Factors influencing the consumption of alcohol and tobacco: a reviw of demand models, Discussion Paper 17 (Centre for Health Economics, University of York).

Gravelle, H.S.E. (1984). Time-series analysis of mortality and unemployment, **Journal of Health Economics 3,** 297-306.

Gravelle, H.S.E. and R. Rees (1981). **Microeconomics**, (Longman, London).

Gray, A. (1982). The production of dental care in the British National Health Service, **Scottish Journal of Political Economy 29**, 59-74.

Gray, A., A. McGuire and P. Stuart (1986). Factor input in NHS hospitals, Discussion Paper 02/86 (Health Economics Research Unit, University of Aberdeen).

Greene, W.H. (1982). Maximum likelihood estimation of stochastic frontier production function models, **Journal of Econometrics 18**, 285-9.

Grossman, M. (1972). **The Demand for Health: A Theoretical and Empirical Investigation**, (NBER, New York).

Grossman, M. (1975). The correlation between health and schooling, in: N.E. Terleckyj, ed., **Household Production and Consumption**, (NBER, New York).

Grossman, M. (1977). A survey of recent research in health economics, **American Economist 21**, 14-20.

Grossman, M. (1982). The demand for health after a decade, **Journal of Health Economics 1**, 1-3.

Grossman, M. and L. Benham (1974). Health, hours and wages, in: M. Perlman, ed., **The Economics of Health and Medical Care**, (Macmillan, London).

Gudex, C. (1986). QALYs and their use by the health service, Discussion Paper 19 (Centre for Health Economics, University of York).

Hadley, J. (1982) **More Medical Care, Better Health?**, (Urban Institute, Washington DC).

Hammond, C.J. (1986). Estimating the statistical cost curve: an application of the stochastic frontier technique, **Applied Economics 18**, 971-84.

Hoskins, M.D. (1982a). The supply of nursing staff to non-psychiatric hospitals in England and Wales, Discussion Paper 23 (Department of Economics, University of Leicester).

Hoskins, M.D. (1982b). The effect of pay changes on cohort survival: a study of the supply of midwifery staff to the National Health Service, Discussion Paper 26 (Department of Economics, University of Leicester).

Hurst, J. (1977). Saving hospital expenditure by reducing inpatient stay, Government Economic Service Occasional Paper 14 (HMSO, London).

Illsley, R. and J. LeGrand (1987). The measurement of inequality in health, Discussion Paper 12 (STICERD Welfare State Programme, London School of Economics).

Jacobs, P. (1974). A survey of economic models of hospitals, **Inquiry 11**, 83-97.

Jensen, G.A. and M.A. Morrisey (1986a). Medical staff specialty mix and hospital production, **Journal of Health Economics 5,** 253-76.

Jensen, G.A. and M.A. Morrisey (1986b). The role of physicians in hospital production, **Review of Economics and Statistics 68,** 432-42.

Johnston, J. (1984). **Econometric Methods,** (McGraw-Hill, London).

Knapp, M. (1984). **The Economics of Social Care,** (Macmillan, London).

Lave, J.R. and L.B. Lave (1970). Economic anaylsis for health service efficiency: a review article, **Applied Economics 1,** 293-305.

Lavers, R.J. (1972). The implicit valuation of forms of hospital treatment, in: M.M. Hauser, ed., **The Economics of Medical Care** (Allen and Unwin, London).

Lavers, R.J. (1983). A model of the demand for presciptions, in: **Actes du Xe Colloque International D'Econometric de la Sante,** (Lyon).

Lavers, R.J. and D.K. Whynes (1978). A production function analysis of English maternity hosptals, **Socioeconomic Planning Sciences 12,** 85-93.

LeGrand, J. (1978). The distribution of public expenditure: the case of health care, **Economica 45,** 125-42.

LeGrand, J. (1982). **The Strategy of Equality: Redistribution and the Social Services,** (Allen and Unwin, London).

LeGrand, J. (1985). Inequalities in health: the human capital approach, Discussion Paper 1 (STICERD Welfare State Programme, London School of Economics).

Lindsay, C. M. (1976). A theory of government enterprise, **Journal of Political Economy 84,** 1061-77.

Lindsay, C.M. (1980). **National Health Issues: The British Experience,** (Roche Laboratories, Nutley).

Lindsay, C.M. and B. Feigenbaum (1984). Rationing by waiting lists, **American Economic Review 74,** 404-17.

McAvinchey, I.D. (1984). Measurement and definition of the link between unemployment and health, **Effective Health Care 1,** 287-94.

McGuire, A. (1985). The theory of the hospital: a review of the models, **Social Science and Medicine 20,** 1177-84.

McGuire, A. and R. Westoby (1983). A production function analysis of acute hospitals, Discussion Paper 04/83 (Health Economics Research Unit, University of Aberdeen).

McGuire, A. and H. Williams (1986). Information theory and Scottish hospital cost functions, Discussion Paper 01/86, (Health Economics Research Unit, University of Aberdeen).

Mann, J.K. and D.E. Yett (1968). The analysis of hospital costs: a review article, **Journal of Business 41,** 191-202.

Maynard, A.K. (1983). The production of health and health care, **Journal of Economic Studies 10,** 31-45.

Maynard, A.K., M. Marinker and D.P. Gray (1986). The doctor, the patient and their contract III - Alternative contracts: are they viable? **British Medical Journal 292,** 1438-40.

Meeusen, W. and J. Van den Broeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed error, **International Economic Review 18,** 435-44.

Migue, J.L. and G. Belanger (1974). Toward a general theory of managerial discretion, **Public Choice 17,** 27-42.

Muurinen, J.M. (1982). An economic model of health behaviour - with empirical applications to Finnish health survey data, unpublished DPhil dissertation (Department of Economics, University of York).

Muurinen, J.M. and J. LeGrand (1984). The economic analysis of inequalities in health, **Social Science and Medicine 20,** 1029-35.

Narendranathan, W., S. Nickell and D. Metcalf (1985). An investigation into the incidence and structure of sickness and unemployment in Britain 1965-75, **Journal of the Royal Statistical Society Series A 148,** 254-67.

Newhouse, J.P. (1981). The demand for medical care services: a retrospect and prospect, in: J. Van der Gaag and M. Perlman, eds., **Health, Economics and Health Economics,** (North-Holland, Amsterdam).

Newhouse, J.P. and L.J. Friedlander (1980). The relationship between medical resources and measures of health: some additional evidence. **Journal of Human Resources 15,** 200-18.

Niskanen, W.A. (1971). **Bureaucracy and Representative Government,** (Aldine, Chicago).

Olowakure, T.O. (1978). Variations in average length of stay among acute hospitals, **Applied Economics 10,** 1-10.

Parkin, D. (1980). Distance as a user cost affecting accessibility to health services, unpublished DPhil dissertation (Department of Economics, University of York).

Parkin, D. and B. Yule (1985). Patient charges and the demand for dental care in Scotland 1962-81, Discussion Paper 04/85, (Health Economics Research Unit, University of Aberdeen).

Pauly, M.V. (1980). **Doctors and Their Workshops: Economic Models of Physician Behaviour,** (University of Chicago Press, Chicago).

Pindyck, R.S. and D.L. Rubinfeld (1981). **Econometric Models and Economic Forecasts,** (McGraw-Hill, London).

Reinhardt, U. (1972). A production function for physician services, **Review of Economics and Statistics 54,** 55-56.

Schmidt, P. (1986). Frontier production functions, **Econometric Reviews 4,** 289-328.

Smith, G. (1983). National Health Service performance indicators, **Public Finance and Accountancy 10,** 17-18.

Spicer, M.W. (1982). The economics of bureaucracy and the British National Health Service, **Milbank Memorial Fund Quarterly 60,** 657-72.

Steele, R. and A.M. Gray (1982). Statistical cost analysis: the hospital case, **Applied Economics 14,** 491-502.

Stern, J. (1983). The relationship between unemployment, morbidity and mortality in Britain, **Population Studies 37,** 61-74.

Stigler, G.J. (1976). The Xistence of X-inefficiency, **American Economic Review 66,** 213-6.

Tatchell, M. (1983). Measuring hospital output: a review of the service-mix and case-mix approaches, **Social Science and Medicine 17,** 871-83.

Townsend, P. and N. Davidson (1982). **Inequalities in Health: The Black Report,** (Penguin Books, Harmondsworth).

Wagstaff, A. (1985a). Time-series analysis of the relationship between unemployment and mortality: a survey of econometric critiques and replications of Brenner's studies, **Social Science and Medicine 21,** 985-96.

Wagstaff, A. (1985b). Unemployment and health: an economic analysis, unpublished DPhil dissertation (Department of Economics, University of York).

Wagstaff, A. (1986a). Unemployment and health: some pitfalls for the unwary, **Health Trends 4,** 79-81.

Wagstaff, A. (1986b). The demand for health: some new empirical evidence, **Journal of Health Economics 5,** 195-233.

Wagstaff, A. (1987). Measuring technical efficiency in the National Health Service: a stochastic frontier analysis, Discussion Paper 30 (Centre for Health Economics, University of York).

Wickens, I. and J. Coles (1985). The ethical imperative of clinical budgeting, Nuffield/York Portfolio 10, (Nuffield Provincial Hospitals Trust, London).

Williams, A. (1977). Health service planning, in: M.J. Artis and A.R. Nobey, eds., **Studies in Modern Economic Analysis,** (Basil Blackwell, Oxford).

Williams, A. (1985). The economics of coronary artery bypass grafting, **British Medical Journal 291,** 326-9.

Wolfe, J.R. (1985). A model of declining health and retirement, **Journal of Political Economy 93,** 1258-67.

Yule, B. and D. Parkin (1985). The demand for dental care: an economic assessment, **Social Science and Medicine 21,** 753-60.